Subject: Social Media Analytics

Module 1: Social Media Analytics: An Overview

SOCIAL MEDIA:

Built on the Web 2.0 philosophy (i.e., give more control to the user over the content), social media is an easy-to-use Internet-based platform that provides users with opportunities to create and exchange content (such as text, videos, audio, and graphics) in a many-to-many context. Social media and Web 2.0 are often use interchangeably, but they can be slightly differentiated. At the core of social media is the Web 2.0 concept, and social media can be considered an application of the Web 2.0 concept.

In other words, social media is realized based on the Web 2.0 concept. One important thing to note is that social media is not limited only to the well-known platforms such as Facebook, Twitter, YouTube, and blogs. We will consider a social media as any online platform (proprietary or purpose built) that enable users to participate, collaborate, create, and share content in a many-to-many context.

CORE CHARACTERISTICS OF SOCIAL MEDIA:

The best way to understand social media is through its core characteristics that set it apart from the conventional medium. All these properties play an important role in creating a collaborative ecosystem.

1. **SOCIAL MEDIA IS MANY-TO-MANY:** Social media enables interaction among the users in a many-to-many fashion. This is unlike conventional technological media such as print, radio, telephone, and television.

2. **SOCIAL MEDIA IS PARTICIPATORY:** Unlike conventional technologies, social media encourages participation and feedback from users. Social media users can participate in online discourse through blogging, comments, tagging, and sharing content.

3. **SOCIAL MEDIA IS USER OWNED:** While social media platforms are provided by corporations (such as Google and Facebook), the content is generated, owned, and controlled by social media users. Without the user-generated contents and active involvement from the users, social media would be empty, boring online spaces.

4. **SOCIAL MEDIA IS CONVERSATIONAL:** It is not only the ease of conversation but also the many-to-many conversation abilities that make social media stand out from the traditional one-to-one or one-to-many medium of interaction. The many-to-many conversation characteristics of social media make it possible for the masses to communication and collaborate in real time.

5. **SOCIAL MEDIA ENABLES OPENNESS:** Social media provides new opportunities to access data and information through Web 2.0 channels.

6. **SOCIAL MEDIA ENABLES MASS COLLABORATION:** Social media channels allow masses to collaborate in a many-to-many fashion to achieve certain shared goals.

7. **SOCIAL MEDIA IS RELATIONSHIP ORIENTED:** Most social media tools allow users to easily establish and maintain social and professional relationships and ties. Some social media tools, such as Facebook, are solely focused on personal relationships, whereas others, such as Twitter, are focused on professional relationships.

8. **SOCIAL MEDIA IS FREE AND EASY TO USE:** Being free and easy to use are two of the reasons that social media has proliferated in such a space.

TYPES OF SOCIAL MEDIA:

Broadly speaking, based on authentication or access mechanisms, social media tools are available in two forms: 1) Internet-based, and 2) smartphone-based.

Internet-based social media platforms are generally accessed through e-mail IDs. Facebook, LinkedIn, Cyworld, and Google+ are examples of Internet-based social media. Note that an Internet-based social media platform can also be accessed through any device connected to the Internet, including a smartphone application (or app for short), but the authentication mechanism is still the same.

Smartphone-based social media platforms are accessed through mobile phone numbers; that is, users can only log in using mobile phone numbers. KaKao Talk, Tender, and 1KM are the popular example of phone-based social network services. These application can only installed and accessed from a phone; for example, in its current form, one cannot use KaKao Talk, 1KM, or Tender through a personal computer. Mobile applications are also an example of mobile-based social media tools.

Below are different social media tools and how businesses can leverage them.

1. SOCIAL NETWORKING SITES: Social networks sites or services (SNS) are types of social media platforms that are solely focused on online social relationships among users. Some examples of SNS include Facebook, Google+, and Cyworld. SNS allow users to build and maintain social relationships among people who share interests, activities, backgrounds, or real-life connections. Most SNS allow "users to
   (1) construct a public or semipublic profile,
   (2) establish links (friendship) and relationships with other SNS users, and
   (3) view and traverse their list of connections and those made by others within the system."

   USING FACEBOOK FOR BUSINESS PURPOSES: Apart from its primary function as an online social network site, Facebook has become an important marketing and outreach channel for all sorts of organizations including governments.

   Fan pages can serve as excellent advertisement and networking channels with your customers. Facebook fan pages are a great way to connect and network with customers. The following questions may bring some clarity and focus to your Facebook fan page efforts.
   ✓ What is the purpose of the Facebook page? And is the purpose aligned with your business goals?

✓ Who will be responsible for handling your Facebook page (e.g., posting information, responding to comments and complaints)?
✓ What should be your Facebook page name?
✓ What information should be posted and what should not be posted?
✓ Do you have a legal mandate to establish an official Facebook page for your organization?
✓ Do you have a plan to collect and analyze feedback generated over your Facebook page?
✓ What are your security measures from possible online risks?

2. CONTENT COMMUNITIES: Content communities, such as YouTube and Flicker, are defined by "a group of people coalescing online around an object of interest held in common. The object can be just about anything for example, photos, videos, links, topic or issue, and is often organized and developed in a way that either includes social network elements or makes them central to the content." The most popular content community site is YouTube.

   USING YOUTUBE FOR BUSINESS PURPOSES: YouTube is a video-sharing website on which users can upload, view, and share videos. An important feature of YouTube is the YouTube channel. A YouTube channel is a public online space (or page) on YouTube. A YouTube channel allows you to upload videos, leave comments, or make playlists. Businesses from around the world use YouTube channels in a variety of ways. For example, it is a great way to advertise, educate customers by uploading training materials, awareness videos, information about your product and services.

   Before configuring a YouTube channel, the some questions raised should be reviewed and answered. Answers to most of the questions should be rooted in your social media strategy
   ✓ What is the purpose of the YouTube channel? (e.g., advertisement promotes awareness, share useful content, provide training etc.) And is the purpose aligned with your business goals?
   ✓ Who will be responsible for handling the channel (e.g., creating and posting videos)?
   ✓ What should be the name of your channel?
   ✓ What type of content should be posted and what should not be posted?
   ✓ Do you have a legal mandate to establish an official YouTube channel for your organization?
   ✓ Do you have a plan to collect and analyze feedback generated over the channel?
   ✓ How will you secure your channel from possible online risks?

3. BLOGS: A blog is a type of online personal space or website where an individual (or organization) posts content (text, images, videos, and links to other sites) and expresses opinions on matters of personal (or organizational) interest on a regular basis. The most popular blogging platforms arehttp://www.wordpress.com and http://www.bloggers.com. Mostly, blogging does not require technical know-how or programming skills, so ordinary users can easily build and manage a professional-looking blog.

   Important features of a blog include:
   Interactivity—Readers have the ability to leave comments in response to a blog post.
   Archives—Blogs provide archives of past blog entries stored in reverse chronological order (that is, the most recent appears first).

Subscription—Internet users can subscribe to blogs. Subscribed users are alerted when new content is posted on the blog.
Focused—Most blogs are focused on a certain area of interest.

USING BLOGS FOR BUSINESS PURPOSES: An official business blog is not just a business diary or journal, but a great way to build a community of readers and receive early and direct feedback on business issues and solicit innovative ideas. As with other platforms, before creating your business blog, review the following questions.

✓ What is the purpose of the official blog? (e.g., advertisement promotes awareness, solicit ideas.) And is the purpose aligned with the business goals?
✓ Who will be responsible for handling the blog (e.g., creating and posting content)?
✓ What should be the name of your blog?
✓ What type of content should be posted and what should not be posted?
✓ Do you have a legal mandate to establish an official blog for your organization?
✓ Do you have a plan to collect and analyze feedback generated over the blog?
✓ How will you secure the blog?

4. MICROBLOGGING: Microblogging is a miniature version of blogging that allows users to exchange/publish brief messages, including text, images, or links to other websites. The most popular microblogging platform is Twitter. Twitter is an online microblogging service that enables users to send and read short messages commonly known as "tweets." A tweet is a text message limited to one hundred forty characters.

Basic Twitter terminologies:

Tweet: A tweet is a one hundred forty-character message posted via Twitter. You can also include links and pictures in a tweet.

Retweet (RT): A retweet is a reposting of someone else's tweet or message. One way to gauge the popularity of your tweets is by measuring retweets. Popular tweets get reposted many times.

Direct messages: Unlike a tweet, which is public and seen by everyone, a direct message is a personal tweet (like e-mail) seen only by the sender and the recipient. However, a direct message can only be sent to people following you.

Following: Following is how you subscribe to other people over Twitter. On Twitter, following someone means that: You are subscribing to their tweets as a follower (their tweets will appear on your Twitter main page). Their updates will appear in your Home tab. That person is able to send you direct messages.

Followers: Followers are people who follow you over Twitter. If someone follows you it means that: They will show up in your followers list. They will see your tweets in their home timeline whenever they log in to Twitter. You can send them direct messages. Research suggests that the number of followers and following strongly correlated, meaning that people who follow

more people get more followers vise versa. One of your objects over Twitter is to increase your flowers.

Mention: When another user includes your username preceded by the @ symbol in a tweet, it is called a "mention." Your Mentions tab (on the Notifications page) collects tweets that mention you by your username so you can keep track of conversations others are having with you. Number of mentions is an indication of influence or popularity.

Hashtags (#): The hashtag (#) symbol is used to mark keywords or topics in a tweet. It is an easy way to categorize messages. Clicking on a hashtagged word in any message shows you all other tweets marked with that keyword.

USING TWITTER FOR BUSINESS PURPOSES: Twitter is a great way to keep your customers informed. Businesses from around the world use Twitter to keep customers informed by disseminating news and information almost in real time. Setting up Twitter is a very simple process. However, it should not be taken lightly, as the Twitter channel will officially represent your organization. Before proceeding to setup, do a little bit of planning. The following question may bring some clarity and focus to your Twitter efforts.

✓ What is the purpose of the Twitter account? And is the purpose aligned with your business goals? Your social media strategy will determine the main purpose of using Twitter or any other social media platform.

✓ Who will be responsible for handling it? Since it is not a one-shot deal, once a social media presence is established, it needs to be sustained and managed properly

✓ What should be your Twitter handler or username? Be thoughtful while creating a handle; think of a name that truly sums up your organization.

✓ What information should be posted and what should not be posted? Your departmental information or communication policy may provide a useful place to start.

✓ Do you have legal mandate to establish a Twitter account?

✓ Do you have a plan to collect and analyze feedback generated over Twitter?

✓ How will you secure your account from online security risks?

5. ONLINE COLLABORATIVE PROJECTS: Wikipedia is an example of online collaborative projects. Online collaborative projects/tools allow people to plan, coordinate, add, control, and monitor content in collaboration with others. At the core of the online collaborative projects is the concept of wiki. A wiki is a type online content management system that allows users to add, modify, or delete content simultaneously in collaboration with others. Famous examples of wiki-based platforms are Wikipedia and wiki-spaces. The concept of wiki was first conceived by Ward Cunningham.

   USING WIKIS FOR BUSINESS PURPOSES: Wikis are a great way to communicate and collaboratively work on projects with other people. A good example of collaborative wiki is http://www.wikipedia.com. Wikipedia has more than thirty million articles in 287 languages written collaboratively by volunteers around the world. However, Wikipedia is just one type of website built on the wiki model. There are several other notable wikis. Google (www.sites.google.com) provides project wikis that can be configured for business purposes.

The questions discussed in earlier sections should be reviewed before configuring a business wiki.

6. FOLKSONOMIES OR TAGGING (E.G., DEL.ICIO.US): The term folksonomy, also known as social tagging, social indexing, and collaborative tagging, is attributed to Thomas Vander Wal. The word was created by fusing folk and taxonomy. In simple words, it is the method of organizing data and content (through tagging) from a user's perspective. For example, del.icio.us, a social bookmarking system, allows users to tag, organize, classify, and share content (web addresses or sites) in their own unique ways. These days, almost all prominent companies (e.g., Facebook and Flicker) also provide tagging services to their users. Since the contents are tagged with useful keywords, social tagging expedites the process of searching and finding relevant content.

7. VIRTUAL WORLDS: Virtual worlds is computer-generated online environments. It can take the form of a three-dimensional (3-D) virtual social world (e.g., Second Life) where people digitally represent themselves in the form of avatars and interact with others through text and voice messaging. It can also take a form of virtual interactive games, such as World of Warcraft. Mostly, the virtual world environment is created by the users themselves. Virtual reality is another dimension of virtual worlds, where real and virtual are fused together. Virtual reality uses computer software and hardware tools to simulate physical presence the virtual world.

8. MOBILE APPS: Mobile apps are becoming an integral part of our lives. Mobile apps are special-purpose tools developed to perform a variety of activities we do every day while on the move, such as communicating, social networking, sharing information, and shopping. Tender and Skout, for example, are designed to facilitate social relations, and Viber is designed to facilitate communication.

9. PURPOSE-BUILT PLATFORMS: Social media is not only limited to the aforementioned types, but any online platform (including purposely built in-house platforms) that enable us to participate, collaborate, create, and share content in a many-to-many context can be called social media. Content can be anything, including information, audio/video, profiles, photographs, text, etc. Organizations are increasingly creating purpose-built social media platforms for inter-organizational collaboration activities. A good example of such a platform is the Enterprise 2.0, (McAfee 2006) which uses social media tools (such as blogs, wikis, and group messaging software) to allow employees, suppliers, and customers to network together and share information.

PURPOSE OF SOCIAL MEDIA ANALYTICS

The main premise of social media analytics is to enable informed and insightful decision making by leveraging social media data (Chen, R.H.L. et al. 2012; Bekmamedova and Shanks 2014). The following are some sample questions that can be answered with social media analytics:

- What are customers using social media saying about our brand or a new product launch?
- Which content posted over social media is resonating more with my customers?

- How can I harness social media data (e.g., tweets and Facebook comments) to improve our product/services?
- Is the social media conversation about our company, product, or service positive, negative, or neutral?
- How can I leverage social media to promote brand awareness?
- Who are our influential social media followers, fans, and friends?
- Who are our influential social media nodes (e.g., people and organizations) and their position in the network?
- Which social media platforms are driving the most traffic to our corporate website?
- Where is the geographical location of our social media customers?
- Which keywords and terms are trending over social media?
- How active is social media in our business and how many people are connected with us?
- Which websites are connected to my corporate website?
- How are my competitors doing on social media?

SOCIAL MEDIA VS. TRADITIONAL BUSINESS ANALYTICS

While the premise of both social media and traditional business analytics is to produce actionable business, they do however slightly differ in scope and nature. Table 1 provides a comparison of social media analytics with conventional business analytics.

The most visible difference between the two comes from the source, type, and nature of data mined. Unlike the traditional business analytics of structured and historical data, social media analytics involves the collection, analysis, and interpretation of semistructured and unstructured social media data to gain an insight into the contemporary issues.
Another visible difference comes from the way the information (i.e., text, photographs, videos, audio, etc.) is created and consumed. Social media data originates from the public Internet and is socialized in nature. The conventional business data is confined within the organizational databases, limitedly shared, and can serves as a source of competitive advantage.

Table 1. Social media vs. conventional business analytics

Social Media Analytics	**Business Analytics**
Semistructured and unstructured data	Structured data
Data is not analytical friendly	Data is analytical friendly
Real-time data	Mostly historical data
Public data	Private data
Stored in third-party databases	Stored in business-owned databases
Boundary-less data (i.e., Boundary within the Internet)	Bound within the business intranet
Data is high volume	Data is medium to high volume
Highly diverse data	Uniform data
Data is widely shared over the Internet	Data is only shared within organizations
More sharing creates greater value/impact	Less sharing creates more value

No business control over data	Tightly controlled by business
Socialized data	Bureaucratic data
Data is informal in nature	Data is formal in nature

SEVEN LAYERS OF SOCIAL MEDIA ANALYTICS

Social media at a minimum has seven layers of data (Figure 2). Each layer carries potentially valuable information and insights that can be harvested for business intelligence purposes. Out of the seven layers, some are visible or easily identifiable (e.g., text and actions) and other are invisible (e.g., social media and hyperlink networks). The following are seven social media layers that will be discussed in detail in the subsequent chapters.

1. Text
2. Networks
3. Actions
4. Hyperlinks
5. Mobile
6. Location
7. Search engines

Figure 2. Seven layers of social media analytics

LAYER ONE: TEXT Social media text analytics deals with the extraction and analysis of business insights from textual elements of social media content, such as comments, tweets, blog posts, and Facebook status updates. Text analytics is mostly used to understand social media users' sentiments or identify emerging themes and topics.

LAYER TWO: NETWORKS Social media network analytics extract, analyze, and interpret personal and professional social networks, for example, Facebook, Friendship Network, and

Twitter. Network analytics seeks to identify influential nodes (e.g., people and organizations) and their position in the network.

LAYER THREE: ACTIONS Social media actions analytics deals with extracting, analyzing, and interpreting the actions performed by social media users, including likes, dislikes, shares, mentions, and endorsement. Actions analytics are mostly used to measure popularity, influence, and prediction in social media.

LAYER FOUR: MOBILE Mobile analytics is the next frontier in the social business landscape. Mobile analytics deals with measuring and optimizing user engagement with mobile applications (or apps for short).

LAYER FIVE: HYPERLINKS Hyperlink analytics is about extracting, analyzing, and interpreting social media hyperlinks (e.g., in-links and out-links). Hyperlink analysis can reveal, for example, Internet traffic patterns and sources of incoming or outgoing traffic to and from a source.

LAYER SIX: LOCATION Location analytics, also known as spatial analysis or geospatial analytics, is concerned with mining and mapping the locations of social media users, contents, and data.

LAYER SEVEN: SEARCH ENGINES Search engines analytics focuses on analyzing historical search data for gaining a valuable insight into a range of areas, including trends analysis, keyword monitoring, search result and advertisement history, and advertisement spending statistics.

TYPES OF SOCIAL MEDIA ANALYTICS

Like any business analytics, social media analytics can take three forms:
1) descriptive analytics, 2) predictive analytics, and 3) prescriptive analytics.

1) DESCRIPTIVE ANALYTICS: Descriptive analytics is mostly focused on gathering and describing social media data in the form of reports, visualizations, and clustering to understand a business problem. Actions analytics (e.g., no. of likes, tweets, and views) and text analytics are examples of descriptive analytics. Social media text (e.g., user comments), for example, can be used to understand users' sentiments or identify emerging trends by clustering themes and topics. Currently, descriptive analytics accounts for the majority of social media analytics.

2) PREDICTIVE ANALYTICS: Predictive analytics involves analyzing large amounts of accumulated social media data to predict a future event. For example, an intention expressed over social media (such as buy, sell, recommend, quit, desire, or wish) can be mined to predict a future event (such as purchase). Or a business manager can predict sales figures based on historical visits (or in-links) to a corporate website. The TweepsMap tool, for example, can help you determine the right time to tweet for maximum alignment with your audience time zone. Or, based on analyzing your social media users' languages, it can suggest if it is time to create a new Twitter account for another language.

3) PRESCRIPTIVE ANALYTICS: While predictive analytics help to predict the future, prescriptive analytics suggest the best action to take when handling a scenario. For example, if you have groups of social media users that display certain patterns of buying behavior, how can you optimize your offering to each group? Like predictive analytics, prescriptive analytics has not yet found its way into social media data.

SOCIAL MEDIA ANALYTICS CYCLE

Social media analytics is a six step irrelative process (involving both the science and art) of mining the desired business insights from social media data (Figure 3). At the center of the analytics are the desired business objectives that will inform each step of the social media analytics journal. Business goals are defined at the initial sage, and the analytics process will continue until the stated business objectives are fully satisfied. To arrive from data to insights, the steps may vary greatly based on the layers of social media mined (and the type of the tool employed). The following are the six general steps, at the highest level of abstraction, that involve both the science and art of achieving business insights from social media data.

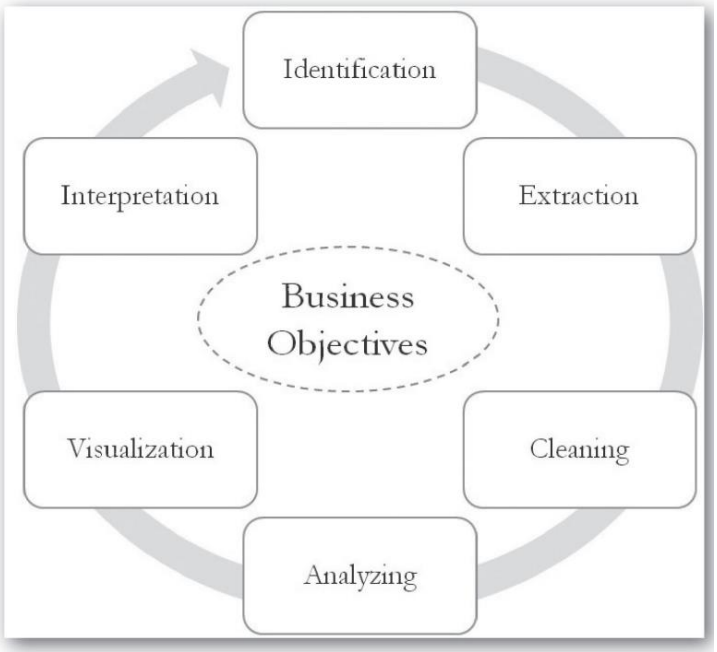

Figure 3. Social media analytics cycle

STEP 1: IDENTIFICATION The identification stage is the art part of social media analytics and is concerned with searching and identifying the right source of information for analytical purposes. Data for analytics will come from business-owned social media platforms. While some data for analytics, will also be harvested from nonofficial social media platforms. The source and type of data to be analyzed should be aligned with business objectives. Framing the right question and knowing what data to analyze is extremely crucial in gaining useful business insights.

STEP 2: EXTRACTION Once a reliable and minable source of data is identified, next comes the science of extraction stage. The type (e.g., text, numerical, or network) and size of data will determine the method and tools suitable for extraction. Small-size numerical information, for example, can be extracted manually (e.g., going through your Facebook fan page and counting likes and copying comments), and a large-scale automated extraction is done through an API (application programming interface).

Two important issues to bear in mind here are the privacy and ethical issues related to mining data from social media platforms. Data extraction practices should not violate a user's privacy and the data extracted should be handled carefully. While all social media platforms have their privacy policies in place, to be on the safe side it is advisable to craft your own social media privacy policy. Your policies should explicitly detail social media ownership in terms of both accounts and activities such as individual and page profiles, platform content, posting activity, data handling and extraction, etc.

STEP 3: CLEANING This step involves removing the unwanted data from the automatically extracted data. Some data may need a lot of cleaning, and others can go into analysis directly. In the case of the text analytics, for example, cleaning, coding, clustering, and filtering may be needed to get rid of irrelevant textual data using natural language processing (NPL). Coding and filtering can be performed by machines (i.e., automated) or can be performed manually by humans. For example, DiscoverText combines both machine learning and human coding techniques to code, cluster, and classify social media data.

STEP 4: ANALYZING At this stage the clean data is analyzed for business insights. Depending on the layer of social media analytics under consideration and the tools and algorithm employed, the steps and approach you take will greatly vary. For example, nodes in a social media network can be clustered and visualized in a variety of ways depending on the algorithm employed. The overall objective at this stage is to extract meaningful insights without the data losing its integrity. While most of the analytics tools will follow you through the step-by-step procedure to analyze your data, having background knowledge and an understanding of the tools and its capabilities is crucial in arriving at the right answers.

STEP 5: VISUALIZATION In addition to numerical results, most of the seven layers of social media analytics will also result in visual outcomes. The science of effective visualization known as visual analytics is becoming an important part of interactive decision making facilitated by solid visualization. Effective visualization is particularly helpful with complex and huge data because it can reveal hidden patterns, relationships, and trends. It is the effective visualization of the results that will demonstrate the value of social media data to top management. Depending on the layer of the analytics, the analysis part will result in relevant visualizations for effective communication of results. Depending on the type of data, different types of visualization are possible, including the following.

Network data (with whom)—network data visualizations can show who is connected to whom. For example, a Twitter following-following network chart can show who is following whom. *Topical data (what)*—topical data visualization is mostly focused on what aspect of a phenomenon is under investigation. A text cloud generated from social media comments can show what topics/themes are occurring more frequently in the discussion.
Temporal data (when)—temporal data visualization slice and dice data with respect to a time horizon and can reveal longitudinal trends, patterns, and relationships hidden in the data. Google trends data, for example, can visually investigate longitudinal search engine trends *Geospatial data (where)*—geospatial data visualization is used to map and locate data, people, and resources.

Other forms of visualizations include trees, hierarchical, multidimensional (chart, graphs, tag clouds), 3-D (dimension), computer simulation, infographics, flows, tables, heat maps, plots, etc.

STEP 6: INTERPRETATION Interpreting and translating analytics results into a meaningful business problem is the art part of social media analytics. This step relies on human judgments to interpret valuable knowledge from the visual data. Meaningful interpretation is particularly important when we are dealing with descriptive analytics that leave room for different interpretations. Having domain knowledge and expertise are crucial in consuming the obtained results correctly. Two strategies or approaches used here can be
1) producing easily consumable analytical results and
2) improving analytics consumption capabilities.

The first approach requires training data scientists and analysts to produce interactive and easy-to-use visual results. And the second strategy focuses on improving management analytics consumption capabilities.

CHALLENGES TO SOCIAL MEDIA ANALYTICS

Social media data is high volume, high velocity, and highly diverse, which, in a sense, is a blessing in terms of the insights it carries; however, analyzing and interpreting it presents several challenges. Analyzing unstructured data requires new metrics, tools, and capabilities, particularly for real-time analytics that most businesses do not possess.

VOLUME AND VELOCITY AS A CHALLENGE: Social media data is large in size and is swiftly generated. Capturing and analyzing millions of records that appear every second is a real challenge. For example, on Twitter, three-hundred forty-two thousand tweets appear every minute, and on Facebook, one million likes are shared every twenty minutes. Capturing all this information may not be feasible. Knowing what to focus on is crucial for narrowing down the scope and size of the data. Luckily, sophisticated tools are being developed to handle high-volume and high-velocity data.

DIVERSITY AS CHALLENGE: Social media users and the content they generate are extremely diverse, multilingual, and vary across time and space. Not every tweet, like, or user is worth looking at. A tweet or mention coming from an influential social media user is more important than a tweet from a noninfluential user. Due to the noisy and diverse nature of social media data, separating important content from noise is challenging and time consuming.

UNSTRUCTUREDNESS AS A CHALLENGE: Unlike the data stored in the corporate databases, which are mostly numbers, social media data is highly unstructured and consists of text, graphics, actions, and relations. Short social media text, such as tweets and comments, has dubious grammatical structure, and is laden with abbreviations, acronyms, and emoticons (a symbol or combination of symbols used to convey emotional expressions in text messages), thus representing a great challenge for extracting business intelligence.

SOCIAL MEDIA ANALYTICS TOOLS

To keep up with the growing need for analyzing the vast amount of data, social media analytical tools are also coming to market at a great pace. Social media analytics tools come in a variety of forms and functionalities. Table 2 lists some example tools with respect to each layer of social media analytics. Aligned with your social media strategy, these tools can be used to measure different layers of social media data.

Table 2. Examples of social media analytics tools with respect its layers

Layer of social media	**Example of tools**
Text	Discovertext Lexalytics Tweet Archivist Twitonomy Netlytic LIWC Voyant
Actions	Lithium Twitonomy Google Analytics SocialMediaMineR
Network	NodeXL UCINET Pajek Netminer Flocker Netlytic Reach Mentionmapp
Mobile	Countly Mixpanel Google Mobile Analytics

Location	Google Fusion Table Tweepsmap Trendsmap Followerwonk Esri Maps Agos
Hyperlinks	Webometrics Analyst VOSON
Research Engines	Google Trends

Subject: Social Media Analytics

Module 2: Social Network Structure, Measures & Visualization

Basics of Social Network Structure:

1. **Nodes:** People in your network are called nodes. Nodes are represented by the circles/dots in the image shown in Figure below.
2. **Edges:** The relationships between people, shown as lines connecting the nodes in Figure below, are called links or edges.

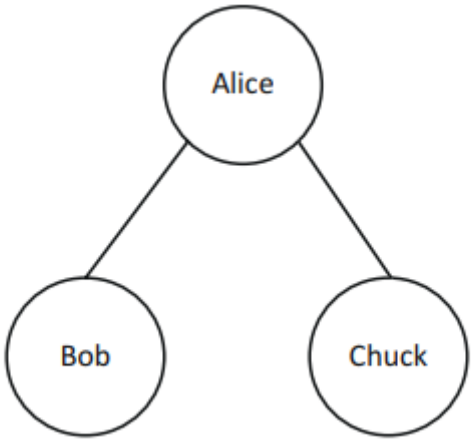

Figure 1: A visualization of a small social network. This shows that Alice has connections to Bob and Chuck, but that Bob and Chuck do not have a connection to one another

3. **Graph:** A group of nodes and edges make up a social network. This is also called a graph or social graph.

Knowing the nodes and edges is all that is needed to analyze a social network. However, edges can have a number of additional features, which can be used in analysis.

Edges can be ***labeled***. The label describes something about the relationship between the people. It could name the relationship (e.g., sister, mother, cousin), or some information about the relationship.

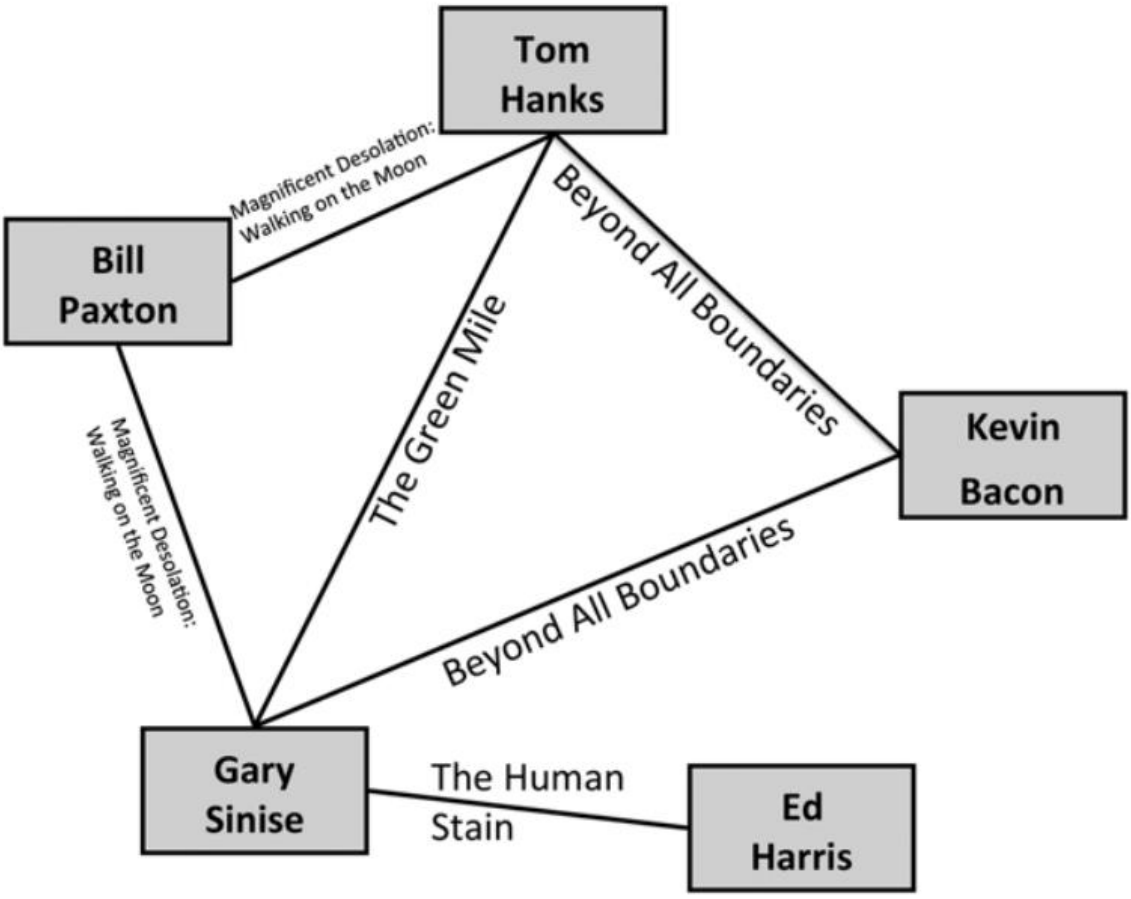

Figure 2: A labeled graph where the edges indicate at least one movie that the actors have been in together

Edges can be ***weighted or valued***. We will use weighted in this book. The weight is a number that indicates numerical information about a relationship. Often, this is the strength of a relationship, but it can come from a variety of sources and indicate many things.

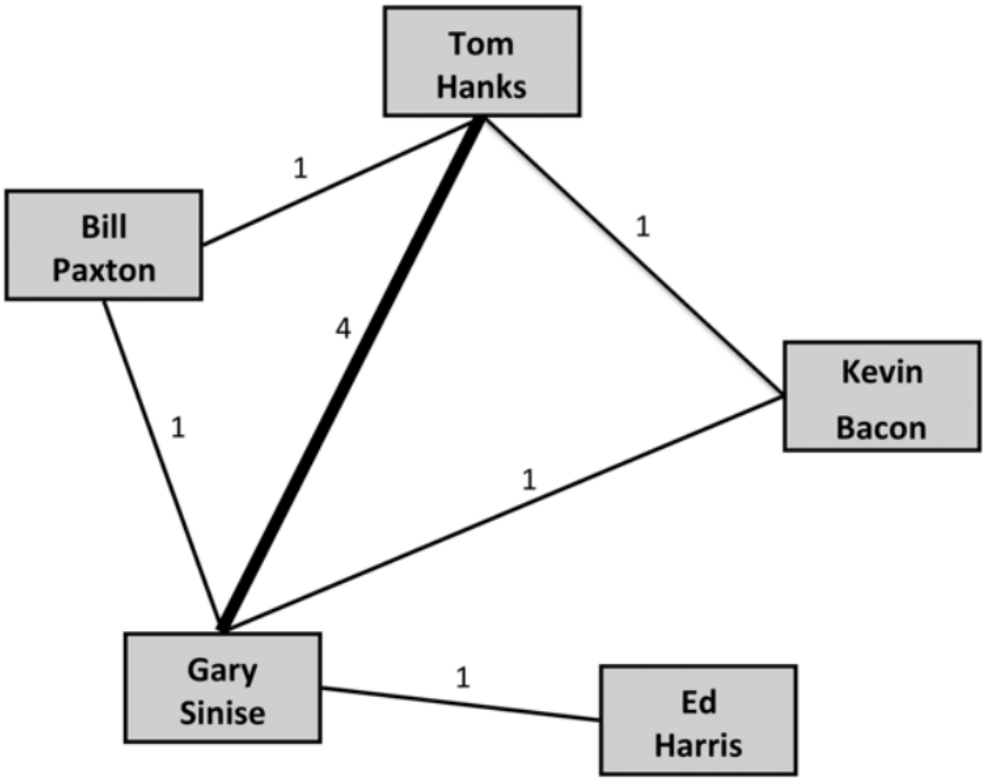

Figure 3: A weighted graph where weights are indicated both as numbers and by the thickness of the edge. In this graph, weight indicates how many movies the actors have been in together.

Edges can also be either ***directed or undirected***. An undirected edge indicates a mutual relationship, whereas a directed edge indicates a relationship that one node has with the other that is not necessarily reciprocated. The type of edge used defines the network as either a directed network or an undirected network.

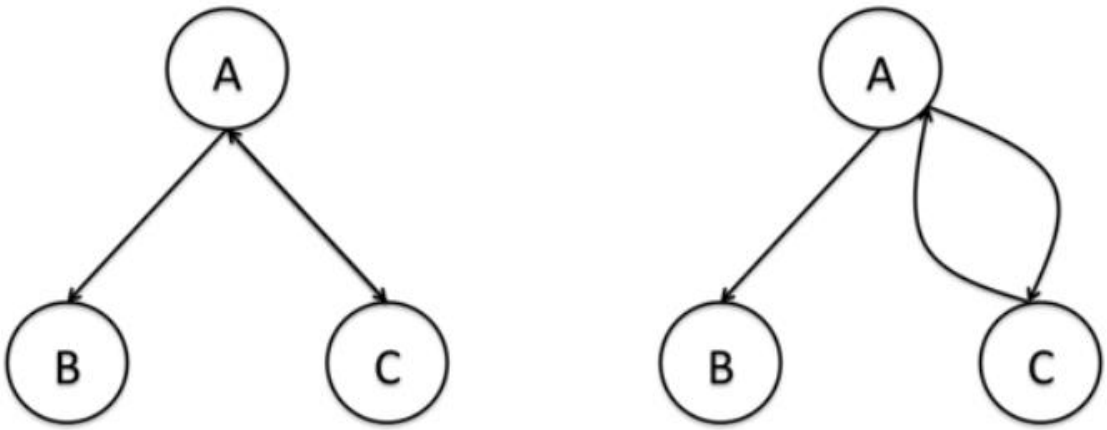

Figure 4: Two ways of drawing a directed network. The edge from A to B is directed only one way. The edge from A to C goes in both directions and can be drawn either as one edge with two arrow heads (left) or as two edges pointing in opposite directions (right).

Representing networks: The example networks we have seen so far are presented as figures with nodes represented as circles or squares and edges as lines that connect them. There are a variety of methods for representing networks. we will focus on text-based representations. These are used as the inputs to many visualization techniques and are also necessary for graphs of most size since they quickly become too large to easily draw as we have in the figures above.

1. **Adjacency lists:** An adjacency list, also called an edge list, is one of the most basic and frequently used representations of a network. Each edge in the network is indicated by listing the pair of nodes that are connected. For example, the adjacency list for the network in figure 3 is as follows:

   Tom Hanks, Bill Paxton,1
   Tom Hanks, Gary Sinise ,4
   Tom Hanks, Kevin Bacon ,1
   Bill Paxton, Gary Sinise ,1
   Gary Sinise, Kevin Bacon ,1
   Gary Sinise, Ed Harris

2. **Adjacency matrix:** An alternative to the adjacency list is an adjacency matrix. In an adjacency matrix, a grid is set up that lists all the nodes on both the X-axis (horizontal) and the Y-axis (vertical). Then, values are filled in to the matrix to indicate if there is or is not an edge between every pair of nodes. Typically, a 0 indicates no edge and a 1 indicates an edge.

The Adjacency Matrix for the Apollo 13 Network

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	1	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	1	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

Notice a couple of things about this matrix.

First, the diagonal is all zeroes because there are no edges between a node and itself in our example. Some networks do allow for self-loops.

Second, the matrix is symmetric. The numbers in the first row are the same as the numbers in the first column. The numbers in the second row are the same as the numbers in the second column. This is because the graph is undirected. Just as in the adjacency list, where the order of pairs in an undirected graph didn't matter.

Notice that the Diagonal, Indicating a Person's Link to Himself, is all 0s

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	1	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	1	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

If we have a directed network, the matrix will not necessarily be symmetric. In the examples we have seen so far, we have been recording a 1 in the matrix to indicate an edge is present, and a 0 when there is no edge. This scheme can be altered to show the weight of an edge as well. To do this, we replace the 1 with the edge weight.

The Adjacency Matrix for the Apollo 13 Network with Edge Weights

	Tom Hanks	Bill Paxton	Gary Sinise	Kevin Bacon	Ed Harris
Tom Hanks	0	1	**4**	1	0
Bill Paxton	1	0	1	0	0
Gary Sinise	**4**	1	0	1	1
Kevin Bacon	1	0	0	0	0
Ed Harris	0	0	1	0	0

XML and standard formats:

In addition to the formats above, a common way to share network data is through standard formats like XML. XML, the eXtensible Markup Language, is the basis for many things on the web, including HTML the language used to write web pages. It is a simple text format designed to be readable by any programming language on any operating system.

An example of how one might represent part of our example network in XML is as follows:

<Person>
<name>Tom Hanks</name>
<connection>Bill Paxton</connection>
<connection>Gary Sinise</connection>
<connection>Kevin Bacon</connection>
</Person>

The text contained between the ,and . signs are tags. There are opening or "start" tags (e.g., ,Person.) and then corresponding end tags that include a leading forward slash. These indicate the end of the section (e.g. ,/Person.). In this snippet of code, we are describing a "Person." The opening tag indicates that our description has started. Between the start and end tags, we list attributes of our person. To do that, we have more tags that describe attributes of the person. This example includes a name (between start and end "name" tags), and the person's

connections. XML can be far more complex than this, but this simple example shows the general structure. Instead of listing pairs of names like we would in adjacency list, connections are represented using XML tags. The benefit of XML is that it is easy to process, and many social network analysis tools are able to read in XML formatted documents to load a social network.

There are a number of standard ways to describe social networks in XML. In the example above, we had a tag for "Person" and tags for "name" and "connection." The XML standards for describing social networks prescribe a set of tag names to use for describing social connections. Examples of these standards include GraphML (the Graph Markup Language) and FOAF1 (Friend Of A Friend).

Basic network structures and properties

Beyond nodes and edges, there are some basic structures that are important to know for describing and understanding networks. These include descriptions of nodes, their connections, and their role in the network.

Subnetworks

So far, we have considered the entire graph or network, looking at how many nodes and edges it has and how to describe them. Often, there are parts of the network that are interesting as well. When we are considering a subset of the nodes and edges in a graph, it is called a subnetwork.

Some of the simplest subnetworks are singletons. These are nodes that have no edges. While these nodes are not very "social," they are still part of a social network. In fact, it is very common to find singletons in online social networks. Often, these represent people who signed up for an account to access some part of the site other than the social networking features, or people who signed up but never actively participated. In Figure 2.6, node A is a singleton because it isn't connected to any other node in the network.

We also are interested in small groups of nodes. When looking at two nodes and their relationship, it is called a dyad, and a group of three nodes is called a triad. Figure 2.6 shows a connected dyad between B and C, and a fully connected triad between D, E, and F. However, we could consider the relationship between A and B. Even though they are not connected, that pair of nodes could also be called a dyad.

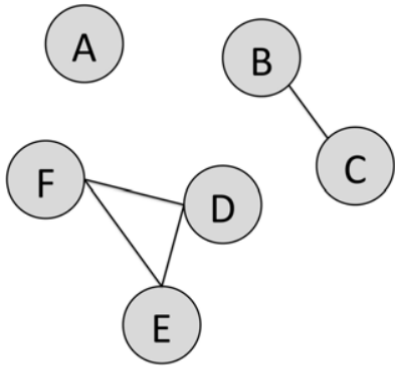

FIGURE 2.6 A social network with a singleton, dyad, and triad

Cliques

Groups of nodes of any size have properties that are interesting. One of particular interest is whether or not all nodes in a group are connected to one another. When this happens, it is called a clique. The term is the same as the one we use to refer to, for example, a group of people who are all strongly connected and tend to talk mostly to one another (e.g., "Alice is part of a clique at school"). For a graph or subgraph to be a clique, every node must be connected to every other. In Figure 2.6, nodes D, E, and F form a clique. However, if the edge from D to E were missing, it would not be a clique.

Clusters

We are also interested in clusters of nodes. In Figure 2.6, we see a group of nodes to the lower right that have many connections between them. This group is not a clique because every node is not connected to every other. For example, node D is not connected to O and F. However, the group is clearly more connected to one another than the graph is as a whole or compared to other subgraphs. While there is no strict definition of a cluster like there is for a clique, we can describe properties of clusters using some network measures, like density, that we will discuss later in this chapter. There are a variety of methods to automatically identify clusters based on the network structure.

Egocentric networks

One of the most important types of subgraphs we will consider is the egocentric network. This is a network we pull out by selecting a node and all of its connections. In Figure 2.6, node D is connected to nodes A, E, B, C, and Q. There are edges from D to each of these nodes and edges between them. When considering egocentric networks, we can choose which of those to include. Consider Figure 2.7. Figure 2.8(a) shows Node D and its edges to its neighbors. Because we are going one step away from D in the network, this is called a degree-1 egocentric network. It only shows us the nodes D is connected to. More frequently, we want to know about the connections between D's neighbors. If we want to see only D's neighbors and their connections, it is called a 1.5-degree egocentric network, shown in Figure 2.8(b). It is 1.5 instead of 2 because we are not going twq full steps away from D in the network. We are going only one step, but then looking at the connections between those

nodes. However, including D in the graph is a bit redundant because we know that D is connected to all of the other nodes. Often, the central node and its edges are excluded and only the

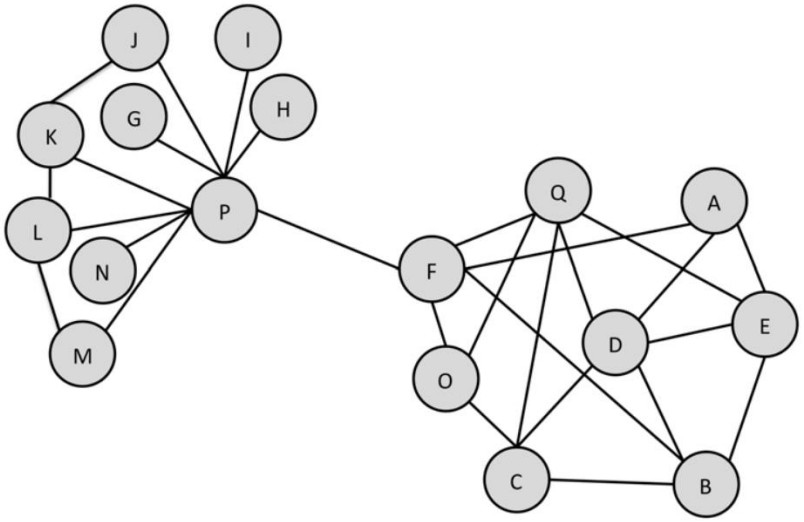

FIGURE 2.7 A sample undirected network

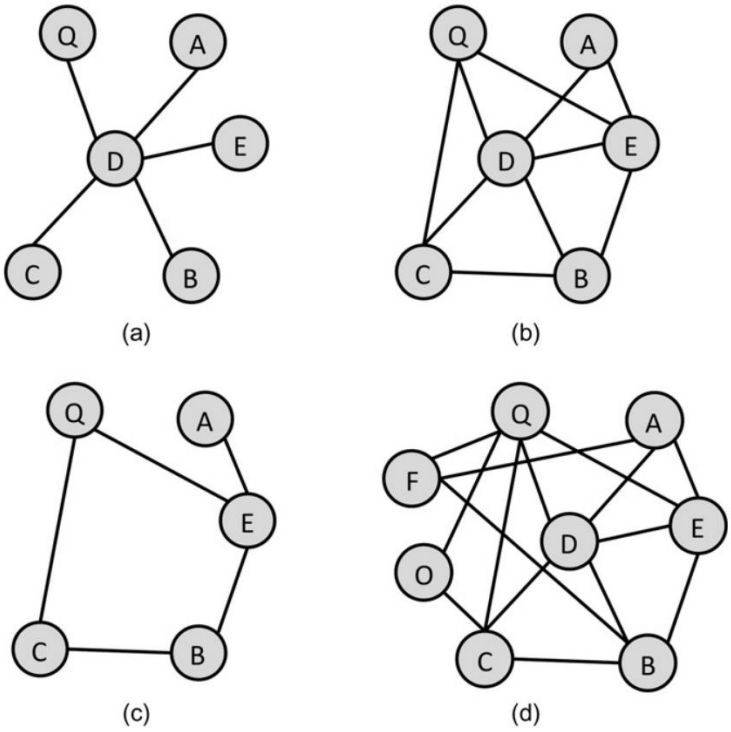

FIGURE 2.8

(a) The 1-degree egocentric network of D, (b) the 1.5-degree egocentric network of D, (c) the 1.5 egocentric network of D with D excluded, and (d) the 2-degree egocentric network of D.

node's neighbors and there connections are considered, as in Figure 2.8(c). This helps make the graph more readable. Egocentric networks can extend out further. Figure 2.8(d) shows the 2-degree egocentric network. It includes all of D's neighbors, their connections to one another, and all of their neighbors. Egocentric networks are used to understand nodes and their role in the network. Egocentric networks are an important tool for network analysis.

Paths and connectedness

The connections between nodes and measures of their closeness are important network characteristics.

Paths: A path is a series of nodes that can be traversed following edges between them. In Figure 2.7, there is a path connecting node M to node C by following the edges from M to P to F to O to C. To determine the length of a path, we count the number of edges in it. The path from M to C has a length of 4 (M-P, P-F, F-O, and O-C). There are longer paths from M to C. For example, we could follow M-L-K-J-P-F-Q-D-C. However, we are typically only interested in the shortest path from one node to another. Note that there may be multiple shortest paths between two nodes. In Figure 2.7, there are two shortest paths from Node F to Node E: F-A-E and F-B-E. Shortest paths will be an important measure we consider in network analysis and are sometimes called geodesic distances.

Connectedness: Paths are used to determine a graph property called connectedness. Two nodes in a graph are called connected if there is a path between them in the network. There does not need to be a direct edge, though that would count. Any path through a series of nodes will work. An entire graph is called ***connected*** if all pairs of nodes are connected.
In an undirected graph, this is relatively straightforward. A path is found by following edges between nodes. In a directed graph, edges may only go in one direction. Thus, while there may be a set of edges that connect two nodes, those edges may not all point in the right direction. If there are edges that can be followed in the correct direction to find a path between every pair of nodes, the directed graph is called ***strongly connected***. If a path cannot be found between all pairs of nodes using the direction of the edges, but paths can be found if the directed edges are treated as undirected, then the graph is called ***weakly connected***.
If a graph is not connected, it may have subgraphs that are connected. These are called ***connected components***. For example, Figure 2.6 includes a three-node connected component, a two-node connected component, and a singleton.

Bridges and hubs: There are two basic concepts that we can use to identify particularly important edges and nodes right off.
The first is a ***bridge***. Intuitively, a bridge is an edge that connects two otherwise separate groups of nodes in the network. Formally, a bridge is an edge that, if removed, will increase the number of connected components in a graph. In Figure 2.7, the edge between nodes P and

F is a bridge because if you take it out, the group of nodes on the right will be totally disconnected from the group of nodes on the left.

Hubs are important nodes rather than edges. They do not have a definition as strict as that of a bridge, but the term is used to refer to the most connected nodes in the network. In Figure 2.7, node P would be a hub because it has many connections to other nodes.

Describing Nodes and Edges:

A node refers to an individual, organization, or group that is represented by a point in the network.

A node is connected to other nodes by lines or edges, which represent relationships or connections between the nodes.

These connections can be based on various factors, such as friendships, familial relationships, shared interests, or professional connections.

The degree of node is the number of edges connected to that nodes. In undirected grapghs. The degree of node is simply the total number of edges connected to it. In directed graphs, there are two measures of degree : in-degree and out-degree.

The in-degree is given by the number of edges coming into the node. In network diagrams, in-degrees are shown as edges with arrows pointing at the node. The out-degree is the number of edges originating from the node going out ward to other nodes. These are shown with arrows pointing away from the node. The sum of the in-degree nad out-degree gives you the total degree for the node.

Determining which nodes are most important or influential is the issue we will discuss in the next section on Centrality.

Why do we need Network Measures?

- Who are the central figures (influential individuals) in the network?
  - **Centrality**

- What interaction patterns are common in friends?
  - **Reciprocity and Transitivity**
  - **Balance and Status**

- Who are the like-minded users and how can we find these similar individuals?
  - **Similarity**

- To answer these and similar questions, one first needs to define measures for quantifying **centrality**, **level of interactions**, and **similarity**, among others.

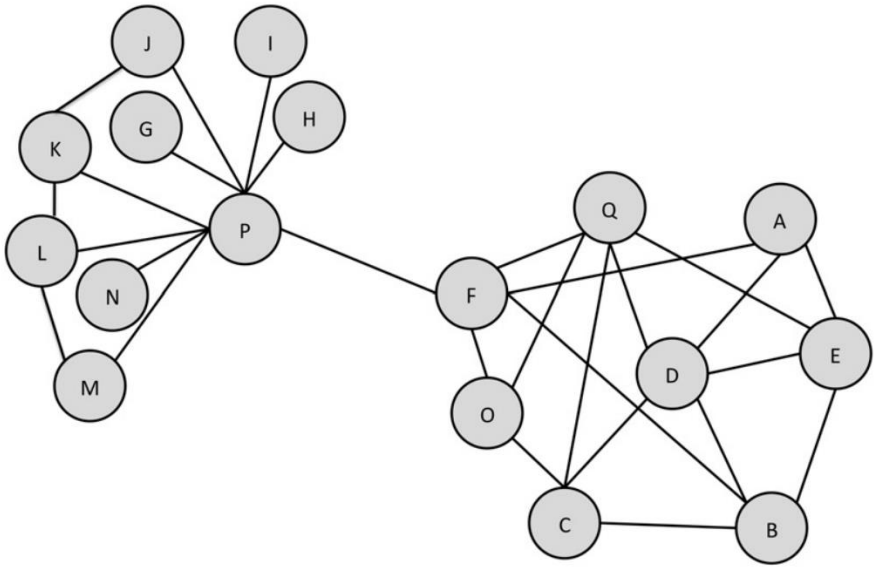

FIGURE 3.1

A sample undirected network.

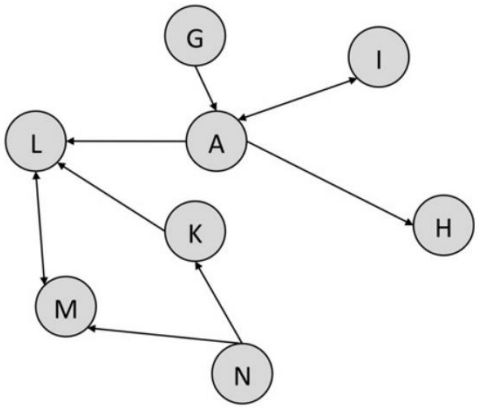

FIGURE 3.2

A sample directed graph.

Centrality

Centrality is one of the core principles of network analysis. It measures how "central" a node is in the network. This is used as an estimate of its importance in the network. However, depending on the application and point of view, what counts as "central" may vary depending on the context. Correspondingly, there are a number of ways to measure centrality of a node. Four types of centrality are considered: degree centrality, closeness centrality, betweenness centrality, and eigenvector centrality.

In network analysis, one or more of these measures may be reported in order to gain a better perspective on the network. A node may appear highly central with one measure but have low centrality with another. That does not mean one measure is incorrect, though; they are simply

different ways of describing nodes. The interpretation of the centrality measures is left to a human analyst.

For all of the centrality measures discussed below, it may be difficult to compare across networks. A very important node in a small network may have centrality measures that would seem unimportant in a larger network. This chapter introduces the basic ways of computing centrality, but they may need to be scaled to facilitate comparisons.

Also, the measures below are calculated for undirected, unweighted graphs. When working with directed or weighted networks, these measures require modification. This has significant implications for how the values are interpreted. Some of these issues will be discussed below with each measure.

Degree centrality

Degree centrality is one of the easiest to calculate. The degree centrality of a node is simply its degree the number of edges it has. The higher the degree, the more central the node is. This can be an effective measure, since many nodes with high degrees also have high centrality by other measures. In Figure 3.1, node P has the highest degree centrality of 9. Meanwhile, node F has a relatively low degree centrality of 5. Many other nodes have that same centrality value or higher (e.g., node D has a degree centrality of 5).

Indeed, as an extreme counterexample, there may be a network with a very large, dense group of nodes that comprise the majority of the graph (this is sometimes called the core of the network), but far out from the core along a chain of low-degree nodes may lie one node that is connected to a large number of nodes with no other connections (this is sometimes said to be on the periphery of the network). This is illustrated in Figure 3.3. Such a node would have high degree centrality, even though it is distant from the core of the network and most of the nodes.

Degree centrality is a good measure of the total connections a node has, but will not necessarily indicate the importance of a node in connecting others or how central it is to the main group.

Closeness centrality

Closeness centrality indicates how close a node is to all other nodes in the network. It is calculated as the average of the shortest path length from the node to every other node in the network. Consider Figure 3.4.

Let's start by computing the average shortest path length of node D. Table 3.1 shows each node and the length of the shortest path from D.

The average of those shortest path lengths is:

$(3 + 2 + 1 + 1 + 2 + 2 + 1) \div 7 = 12 \div 7 = 1.71$

Note that we divide by 7 because there are seven other nodes.

Now repeat this for node A. This is shown in Table 3.2.

Here, the average shortest path length is:

$(1 + 2 + 3 + 4 + 5 + 5 + 4) \div 7 = 24 \div 7 = 3.43$

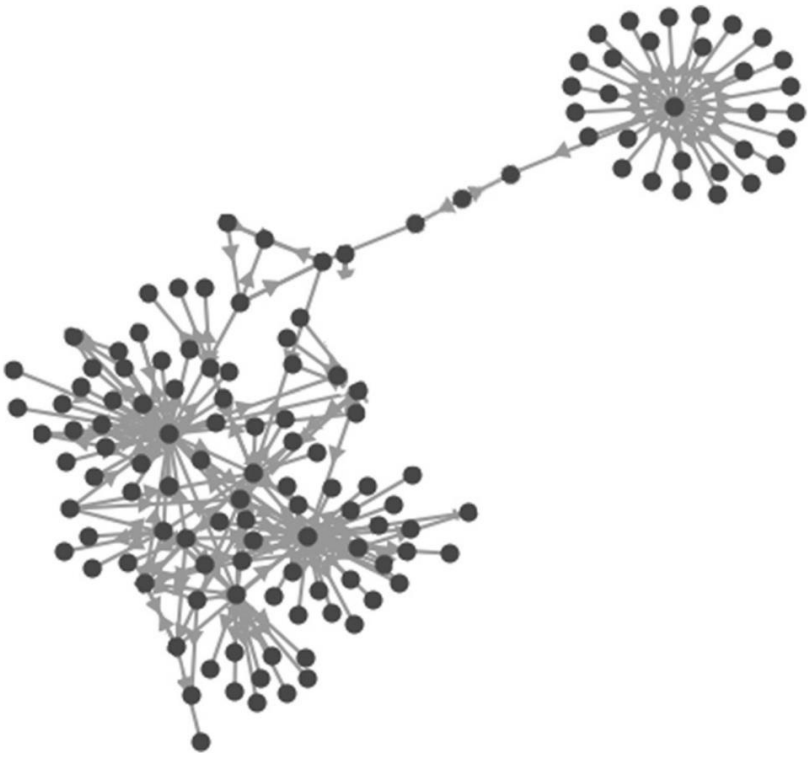

FIGURE 3.3

The node at the center of the cluster in the upper right would have a high degree centrality, even though it is far from the dense center of the network.

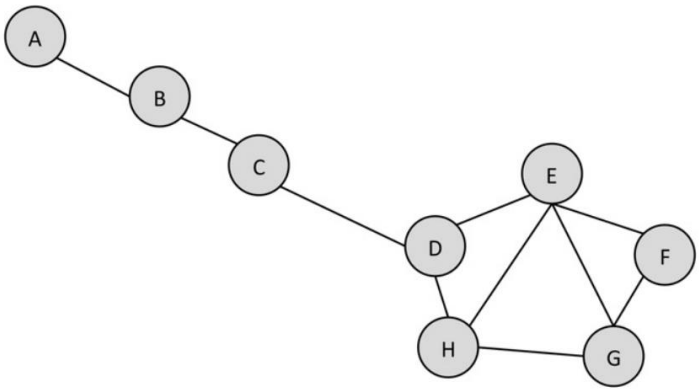

FIGURE 3.4

A sample network.

In the case of closeness centrality, or average shortest path length, lower values indicate more central nodes. Thus, since node D's closeness centrality is 1.71 and node A's is 3.43, node D is more central by this measure.

The benefits of closeness centrality are that it indicates nodes as more central if they are closer to most of the nodes in the graph. This strongly corresponds to visual centrality a node that would appear toward the center of a graph when we draw it usually has a high closeness centrality

Table 3.1 The Shortest Path Lengths from D to each Other Node in the Network

Node	Shortest Path from D
A	3 (D C B A)
B	2
C	1
E	1
F	2
G	2
H	1

Table 3.2 The Shortest Path Length from node A to Every Other Node in the Network

Node	Shortest Path from A
B	1
C	2
D	3
E	4
F	5
G	5
H	4

Betweenness centrality

Betweenness centrality measures how important a node is to the shortest paths through the network. To compute betweenness for a node N, we select a pair of nodes and find all the shortest paths between those nodes. Then we compute the fraction of those shortest paths that include node N. If there were five shortest paths between a pair of nodes, and three of them went through node N, then the fraction would be 3÷5 = 0.6. We repeat this process for every pair of nodes in the network. We then add up the fractions we computed, and this is the betweenness centrality for node N.

For example, consider Figure 3.4. Let's compute betweenness centrality for node B. There are 10 pairs of nodes to consider: AC, AD, AE, AF, CD, CE, CF, DE, DF, and EF. Without counting, we know that 100% of the shortest paths from A to every other node in the network go through B, since A can't reach the rest of the network without B. Thus, the fractions for AC, AD, AE, and AF are all 1.

From C to D, there are two shortest paths: one through B and one through E. Thus, 1 ÷ 2 = 0.5 go through B. The same is true for the shortest path from D to C. For the remaining pairs CE, CF, DE, DF, and EF no shortest paths go through B. Thus, the fraction for all of these is zero. Now we can calculate the betweenness for B:

4 X 1 (A to all others) + 0.5 (DC) + 0.5 (CD) + 5 X 0 (all remaining pairs) = 4 + 0.5 + 0.5 + 0 = 5

In contrast, the betweenness centrality of A is zero, since no shortest paths between D, C, D, E, and F go through A.

Betweenness centrality is one of the most frequently used centrality measures. It captures how important a node is in the flow of information from one part of the network to another.

In directed networks, betweenness can have several meanings. A user with high betweenness may be followed by many others who don't follow the same people as the user. This would indicate that the user is well-followed. Alternatively, the user may have fewer followers, but connect them to many accounts that are otherwise distant. This would indicate that the user is a reader of many people. Understanding the direction of the edges for a node is important to understand the meaning of centrality.

Eigenvector centrality

Eigenvector centrality measures a node's importance while giving consideration to the importance of its neighbors. For example, a node with 300 relatively unpopular friends on Facebook would have lower eigenvector centrality than someone with 300 very popular friends (like Barak Obama). It is sometimes used to measure a node's influence in the network. It is determined by performing a matrix calculation to determine what is called the *principal eigenvector* using the adjacency matrix. The mathematics here are more complicated than this book will cover, but the principles of only is it used to determine

influence in social networks, but a variant of eigenvector centrality is at the core of Google's PageRank algorithm, which they use to rank web pages.

The main principle is that links from important nodes (as measured by degree centrality) are worth more than links from unimportant nodes. All nodes start off equal, but as the computation progresses, nodes with more edges start gaining importance. Their importance propagates out to the nodes to which they are connected. After re-computing many times, the values stabilize, resulting in the final values for eigenvector centrality.

Most network analysis software packages will compute eigenvector centrality (and most other centrality measures as well), so it is not necessary to learn the intricacies of computing eigenvectors. However, understanding the general principles behind the measure is useful to decide when it is the right measure to use in analysis.

Describing networks

A number of measures can be used to describe the structure of a network as a whole. As discussed above, density is one of these. Density the number of edges in the graph divided by the number of possible edges is one of the most common ways of describing a network. However, other statistics provide different insights into network structure.

Degree distribution

Degree is used to describe individual nodes. To get an idea of the degree for all the nodes in the network, we can build the *degree distribution*. This shows how many nodes have each possible degree.

To create a degree distribution, calculate the degree for each node in the network. Table 3.3 shows the degrees for each node in the graph shown in Figure 3.1. The next step is to count how many nodes have each degree. This is totalled for each degree, including those for which there are no nodes with that count. Table 3.4 shows the node count for each degree in this network.

The most common way to show a degree distribution is in a bar graph. The x-axis has the degrees in ascending order, and the Y-axis indicates how many nodes have a given-degree. For the data in Table 3.4, we would make a bar graph as shown in Figure 3.5.

Table 3.3 Degrees for each Node Shown in Figure 3.1

Node	Degree
A	3
B	4
C	4
D	5
E	4
F	4
G	1
H	1
I	1
J	2
K	3
L	3
M	2
N	1
O	2
P	9
Q	5

Table 3.4 The Degree Distribution for the Network in Figure 3.1. The First Column Shows the Degree, and the Second Column Shows How Many Nodes have that Degree

Degree	Number of Nodes
1	4
2	3
3	3
4	4
5	2
6	0
7	0
8	0
9	1

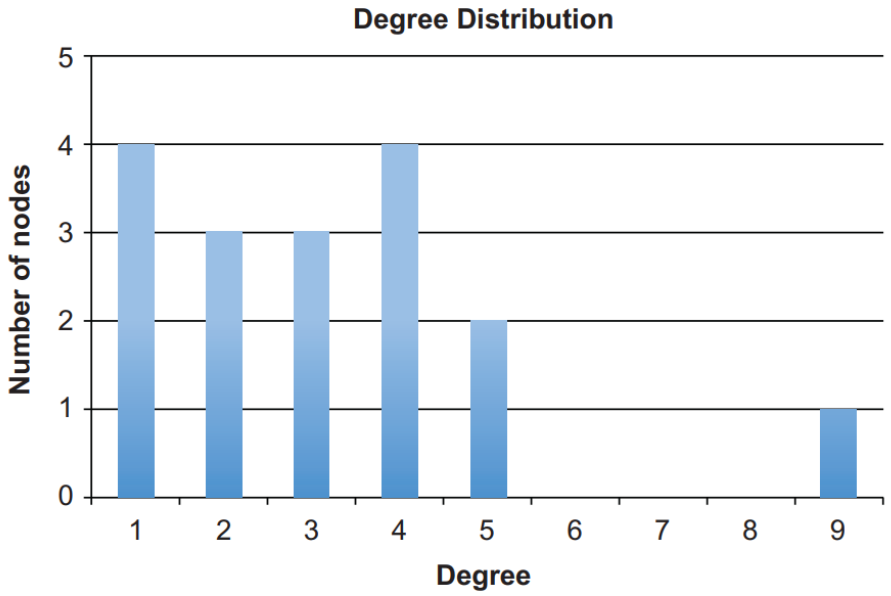

FIGURE 3.5

The degree distribution for the graph shown in Figure 3.1.

Density

Calculating density: A node's connections say a lot about its role in the network. This goes well beyond the degree of a single node or the degrees of all nodes in the network.

Another way to understand both individual nodes and the network as a whole is by studying *density*. Density describes how connected a network is. More formally, it is a statistic comparing the number of edges that exist in a network to the number of edges that could possibly exist. Consider the following two networks, which both have the same number of nodes. Network (a) has very few edges while network (b) has numerous edges among the same number of nodes. Therefore, network (b) has higher density.

There is a formula to calculate density:

number of edges ÷ number of possible edges

The number of edges is something we can count in the network. The number of possible edges could also be counted by looking at each node and counting each of the other nodes that it could connect to. However, there is a simple formula for computing the number of possible edges as well.

First, consider the intuition behind the formula. If there are eight nodes in a network (as there are in the networks in Figure 3.6) each node can connect to seven other nodes. Node A can connect to B, C, D, E, F, G, and H. Node B can connect to A, C, D, E, F, G, and H. This scenario is sometimes known as the *handshake problem* if a person comes into a room, how many people can he or she shake hands with? So if there are eight nodes in a network, and each node can connect to (shake hands with) seven others, then there are 8 X 7 = 56 possible

edges. For a network with n nodes, we can generally say that there are n X (n - 1) edges. Each node can connect with every other node, excluding itself (hence the minus 1).

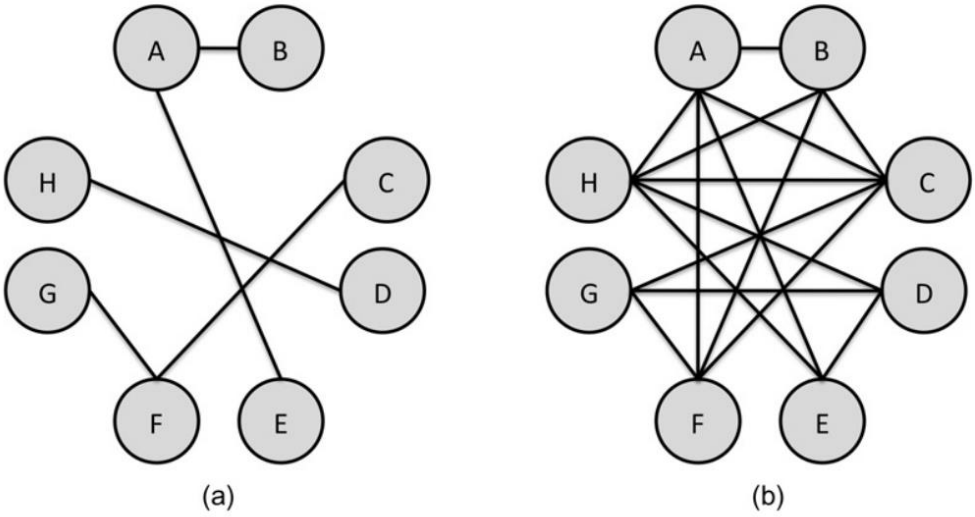

FIGURE 3.6

Network (a) on the left has fewer edges than network (b) on the right. Since they both have the same number of nodes and thus the same number of possible edges, network (b) is more dense.

However, it is not quite that simple. In this example, node A can connect to B and others, and node B can connect to A and others. Since each node can connect to 7 others, each connection has been counted twice. The connection from A to B is counted, as is the edge from B to A. In directed networks, this is fine there are indeed two possible edges between A and B.

But in undirected networks (like the one in Figure 3.4), there can be only one edge between two nodes. Since the formula counts every node twice, simply divide by 2 to count the number of possible edges only once.

Thus, for **directed networks**, the number of possible edges in a graph with n nodes is:

$$n \text{ X } (n - 1)$$

In **undirected networks**, the number of possible edges is:

$$\frac{n*(n - 1)}{2}$$

Now these formulas can be used to calculate density. In a directed network with *n* nodes and *e* edges, the formula for density is:

$$\frac{e}{n*(n - 1)}$$

In an undirected network with *n* nodes and *e* edges, the density formula is:

$$\frac{e}{n*(n - 1)/2}$$

We can use density to describe a network as a whole. Consider the networks in Figure 3.6. Both have eight nodes. Network (a) has five edges. Since it is an undirected network, the density is
(5/(8*(8 – 1))/2) or 5 ÷ 28 = 0.179. Network (b) has 16 edges, so the density is 16 ÷ 28 = 0.571. Note that the density is higher for network (b), meaning it's denser.

A network with no edges would have a density of 0 (because the numerator in our equation would be 0, regardless of how many nodes there are). On the other hand, the densest possible network would be a network where all possible edges exist a clique. As we just learned, the number of possible edges is the denominator of the density formula. In a clique, then, the numerator and denominator will be the same, so the density will be 1. This illustrates that density is always between 0 and 1, where 0 is the lowest possible density and 1 is the highest.

Density in egocentric networks

Density is a common way to compare networks. But it is even more commonly used to compare *subnetworks* especially egocentric networks. Computing the density of each node's egocentric network gives us a way to compare nodes. Some will have dense egocentric networks, which means a lot of their friends know one another. Others will have sparse egocentric networks, and thus we know their connections often do not know one another. The density of an egocentric network is sometimes referred to as the *local clustering coefficient*.
To compute the density of an egocentric network, we use the 1.5-diameter network: We consider the node's connections and all the connections between those nodes.
For this calculation, the ego-node will be excluded from its egocentric network because the density of interest is that of the connections between the node's friends.
As an example, recall the network (b) from Figure 3.6. Node A is connected to nodes B, E, and H. To get the 1.5-diameter egocentric network, we will look at only nodes B, E, and H and the connections between them. This is shown in Figure 3.7(a). There are three nodes, so the number of possible edges is 3 X 2 ÷ 2 = 3.
Possible edges are from B to H and E (2) and from E to H (1) a total of three. There are two edges in the network from H to B and H to E. Thus, the density is 2 (the number of actual edges) 4 3 (the number of possible edges): 2 ÷ 3 = 0.667.
The density of Node B's egocentric network can be computed from the 1.5-diameter egocentric network shown in Figure 3.7 (b). There are four nodes, so the number of possible edges is 4 X 3 ÷ 2 = 6. In the network, there are five edges (from A to H, F, and C, and additionally from C to F and H). So, the density of B's egocentric network is 5 ÷ 6 = 0.833.
Thus, B's egocentric network (0.833) is more dense than A's (0.667). This is a common way to compare nodes in a network. However, having a higher egocentric network density does not necessarily mean a node is more "popular" or important. A node with a high degree (connections to many other nodes) will usually have a lower density. This follows the same logic we discussed above when comparing the density of small networks versus large networks. As the number of nodes in an egocentric network increases, the number of possible edges increases at that rate squared. Thus, more popular nodes tend to have lower densities.

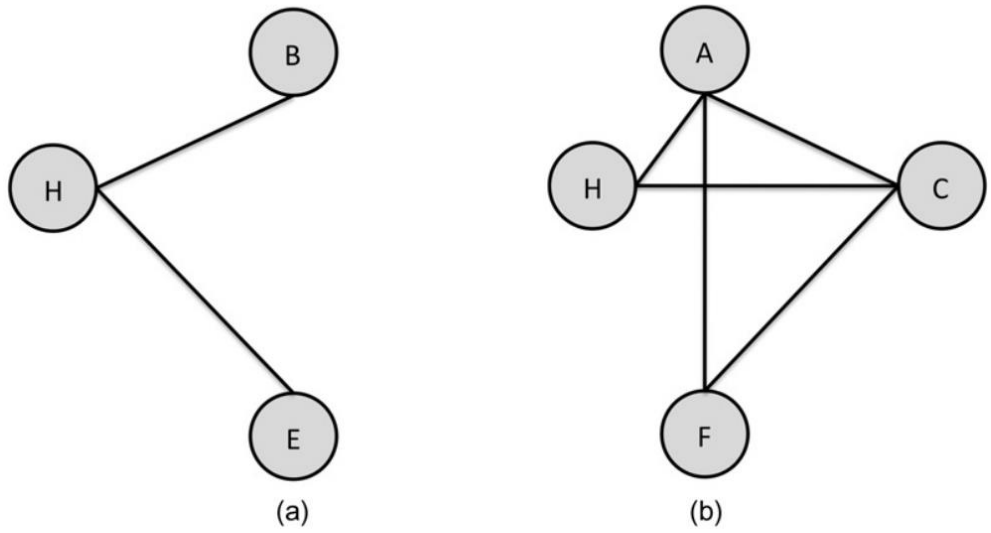

FIGURE 3.7

The 1.5-diameter egocentric networks for nodes A (a) and B (b) from Figure 3.2.

Connectivity

Density measures the percentage of possible edges in a graph. Connectivity, also known as *cohesion*, measures how those edges are distributed. *Connectivity* is a count of the minimum number of nodes that would have to be removed before the graph becomes disconnected; that is, there is no longer a path from each node to every other node.

In Figure 3.4, the connectivity is 1 because removing node B, C, or D would disconnect the graph. Since removing any one of those nodes disconnects the graph, the connectivity is 1. In Figure 3.8, the connectivity is 2. Removing any one node would not break the graph into two parts, but there are several options for removing two nodes that would. For example, removing nodes E and F would separate G from the rest of the graph. If we removed B and D instead, node A would become separated.

Centralization

Centrality is an important way to understand the role of a node in the network and to compare nodes. *Centralization* uses the distribution of a centrality measure to understand the network as a whole. Any of the centrality measures presented above can be used, but only one is used at a time when computing centralization. If one node has extremely high centrality while most other nodes have low centrality, the centralization of the graph is high. If centrality is more evenly distributed, then the centralization of the network is low.

Centralization of power is an often-used concept and phrase, which relates very closely to centralization in a graph. For example, betweenness centrality can represent the control one node has in the ability of others to communicate. If many messages must pass through a particular node along their shortest paths, that node has the power to stop or pass on information. If a few nodes have very high betweenness, we can say that the power is centralized in those nodes.

Centralization is computed by looking at the sum of the differences in centrality between the most central node and every other node in the network, and dividing this by the maximum possible difference in centrality that could exist in the graph (Freeman, 1979). Since there are different centrality measures (e.g., betweenness, closeness, etc.), there are different centralization measures for a graph. But the basic formula is the same, and different centrality measures can be substituted.

Network Visualization

Humans are wired to find patterns visually. We have natural abilities to see anomalies, patterns, clusters, and changes and we can recognize many of these things without consciously looking for them.

Consider Figure 4.1. Without any instructions on what to look for, and without thinking, you can immediately see the circle in the second row standing out from the pattern of squares.

Similarly, in Figure 4.2, it is easy to see the single outlier point that stands apart from the pattern of values in the chart.

And even in graphs, patterns are easy to see. Consider Figure 4.3.

Even knowing nothing about social network analysis, one can see that node a has many neighbors, there is a tight cluster of nodes in the lower right, and there is a long chain from node b running out to node b4.

In visual data patterns can be recognized that may otherwise be difficult to see in lists of numbers, adjacency lists, or other textual representations of data.

Information visualization deals with the presentation of data in visual format. The data may be numeric, categorical, network data (like social networks), text, and other types. Good information visualization supports users in better understanding the data they are seeing.

The goal of information visualization is to take advantage of humans' natural abilities to see patterns, anomalies, relationships, and features in visual data. Visualization provides an overview of complex data. From there, people can identify features of interest, refocus attention on those features, and explore more. Visualizations are a qualitative way to begin understanding data. From there, quantitative experiments or analysis can follow to explain any insights. Graph visualizations apply all these lessons to looking at the structure of networks.

Future chapters will feature the use of graph visualization for understanding networks in many contexts. This chapter will specifically focus on types of network visualizations and the features that are used to help highlight interesting features.

Graph layout

Every network is made up of nodes and edges. How they are laid out is critical to what an observer is able to understand about a network. There are many types of

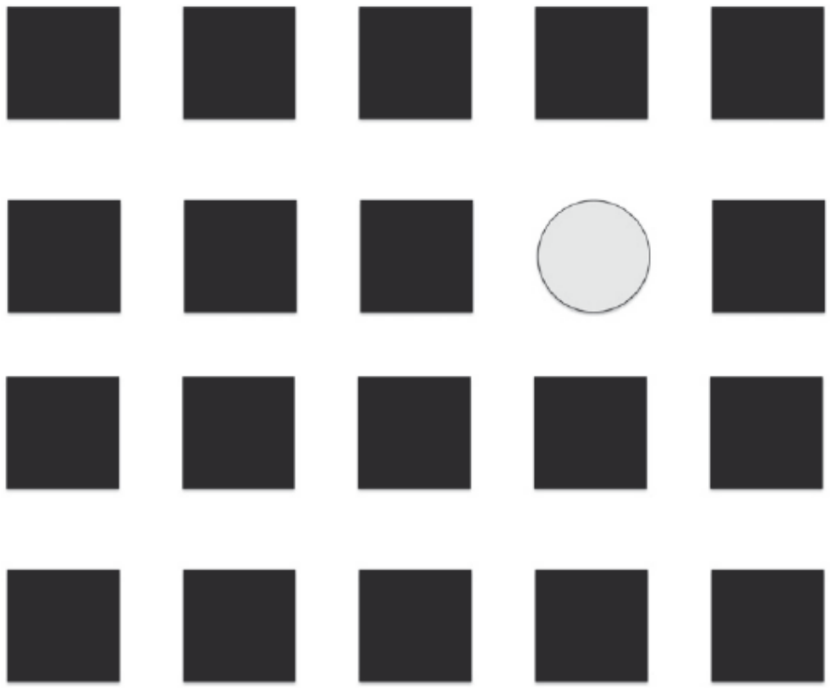

FIGURE 4.1

Without conscious analysis, it is easy to pick out the circle as an anomaly in the pattern.

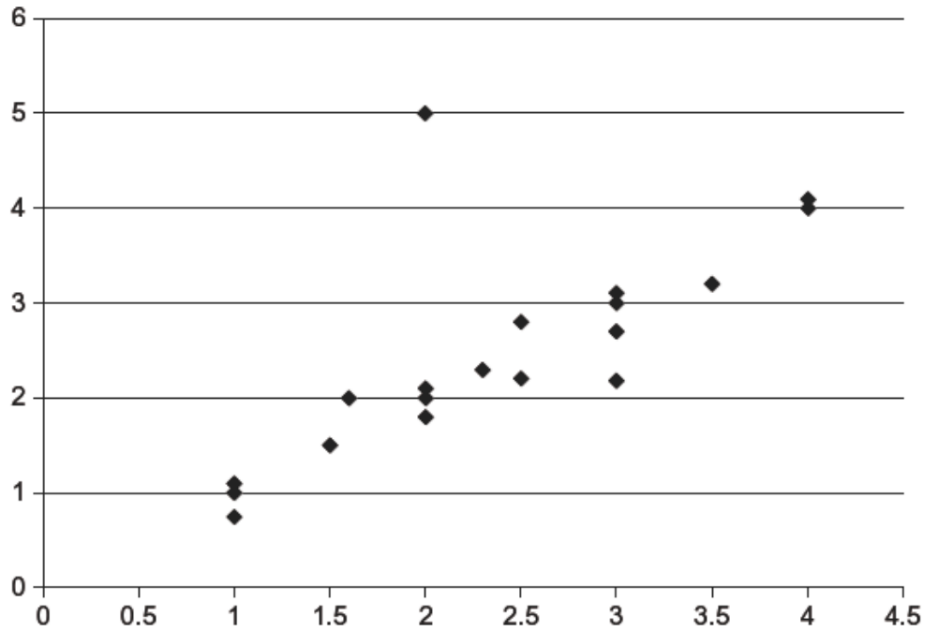

FIGURE 4.2

A single outlier point at value 2 on the x-axis is easy to see separated from the pattern of values.

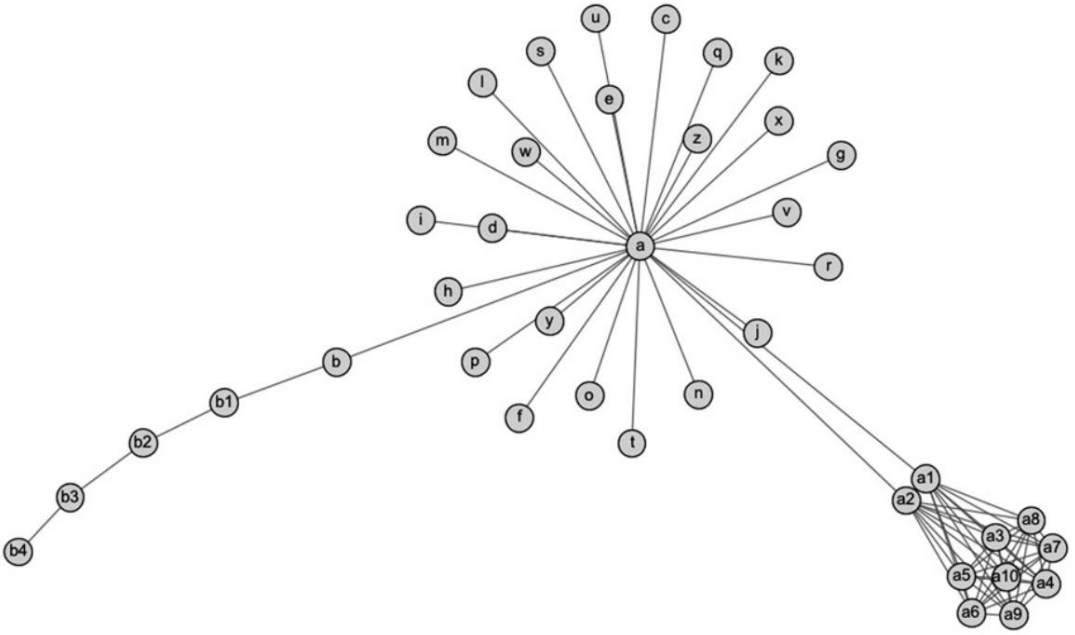

FIGURE 4.3

A sample network visualization.

layout algorithms that position the nodes and edges in different ways when visualizing a network. What makes a "good" layout is not always clear. It depends on what the analyst wants to find, what type of network is being viewed, and what its features are. However, researchers have presented some general guidelines that make network visualizations easier to work with (Dunne and Shneiderman, 2009):

1. Every node is visible.
2. For every node you can count its degree.
3. For every link you can follow it from source to destination.
4. Clusters and outliers are identifiable.

This section presents a few of the most common network layout algorithms.

Random layout

Often, when loading data into a visualization tool, the nodes are placed randomly. This is called a random layout, and it often does not provide much insight into the structure of the network. Figure 4.4 shows the same network from Figure 4.3 presented in a random layout.
We may be able to tell that node a has a high degree in this network, but the clusters and other patters are not at all clear from the random layout.

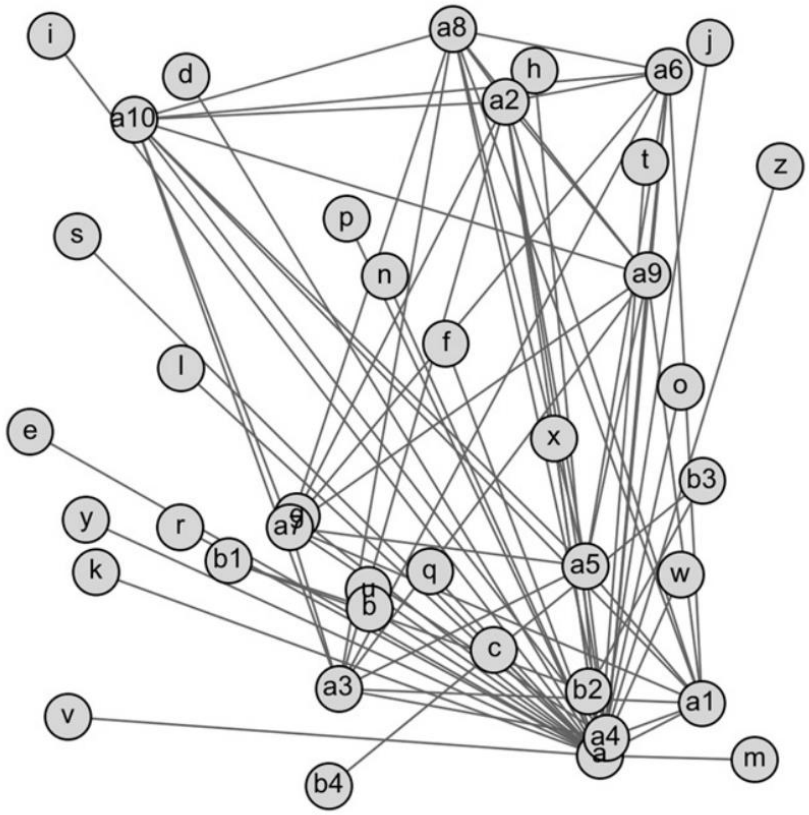

FIGURE 4.4

A random layout of the graph shown in Figure 4.3.

Circular layout

Circular layouts place all the nodes in a circle and then add edges between them. Some circular layouts place nodes closer to one another when they are more closely connected. In Figure 4.5, the cluster of nodes labelled a1 through a10 is clearer because of the density of edges in that section of the graph. The chain of nodes from b through b4 is also present, though the edges around the circle are a bit harder to pick up visually than in the Figure 4.3 layout.

A circular layout places nodes in structured positions and then adds edges between connected pairs. Another way to do this is to place nodes in a grid.

Grid layout

Figure 4.6 shows an example of a grid layout for the same graph in Figures 4.34.5. Note that the degree of node a is clearly high, the cluster of nodes a1 through a10 is obvious, and the chain of nodes b through b4 is clear across the top.

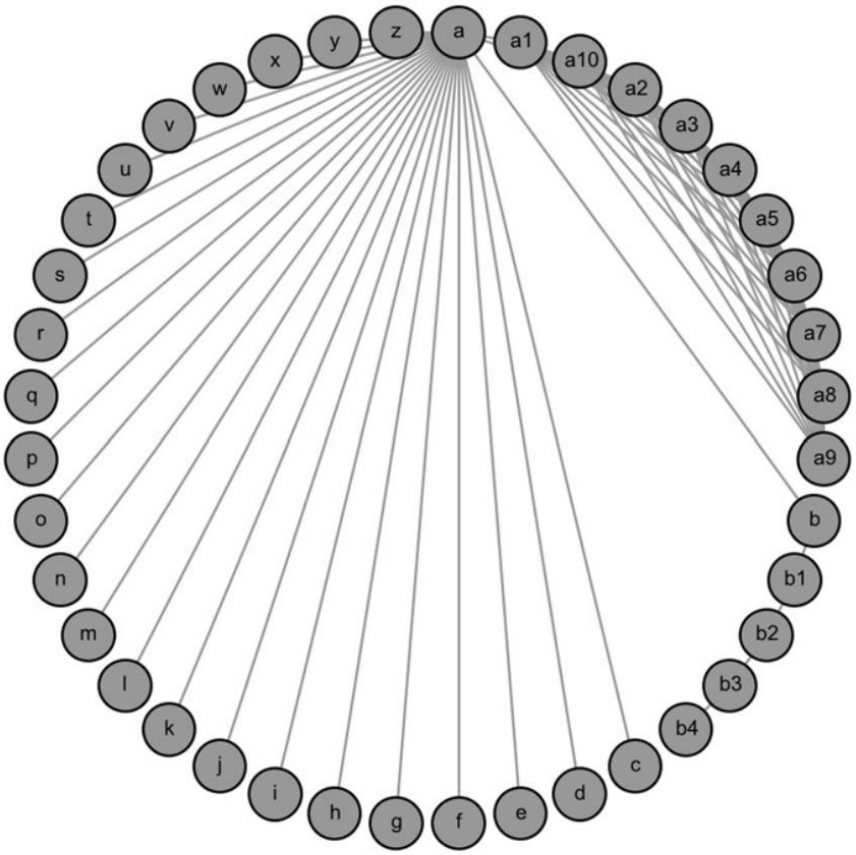

FIGURE 4.5

A circular graph layout for the same graph shown in Figure 4.3.

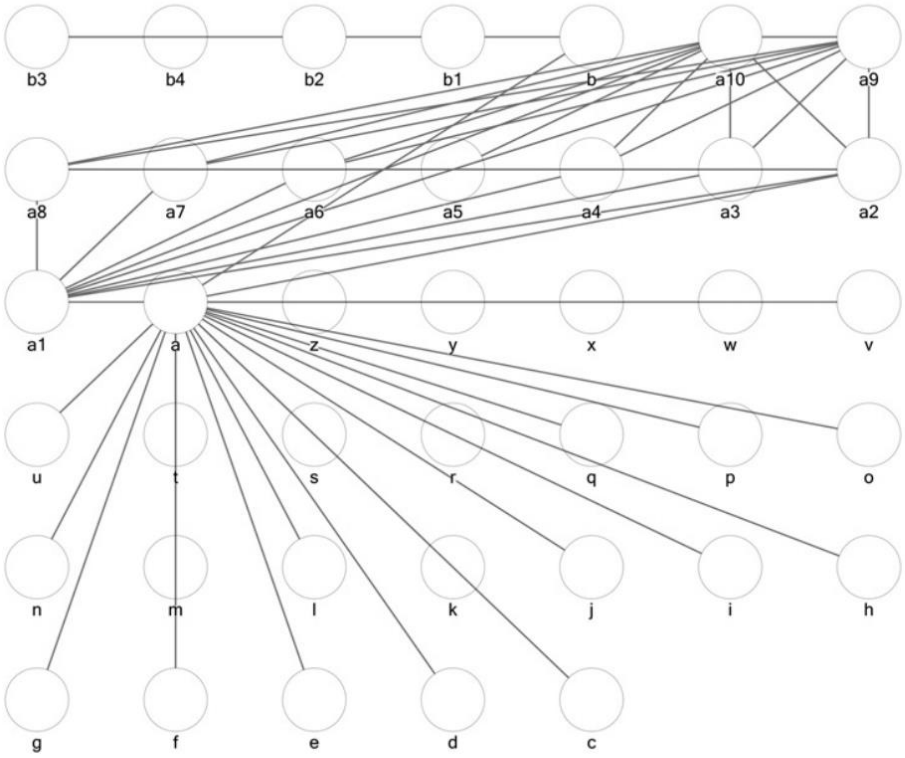

FIGURE 4.6

A grid layout of the modes in the sample graph.

Force-directed layout

Most graphs are not laid out randomly or in one of these formats with a predetermined structure. Instead, the layout is dynamic and determined by the connections between the nodes. Those nodes that are more closely connected are laid out close to one another, and those that are distant are shown further apart.

Figure 4.3 uses an algorithm that does this. This type of layout is generally called *force directed*. Nodes and edges are treated as a physical system, and a simulation of that system is applied to determine a final layout. For example, nodes may be treated as objects, and edges may be treated as springs that apply equal force. The nodes are randomly laid out, connected by springs for edges, and then a simulation of how the springs would physically behave determines the final position of nodes and edges. A cluster of nodes with many connections will be close together, because pulling any node away pulls on many springs that want to keep it close. Nodes with little or no connection are not attracted to one another. Similar approaches to layout that rely on physical simulation include simulated annealing or treating the nodes like charged particles.

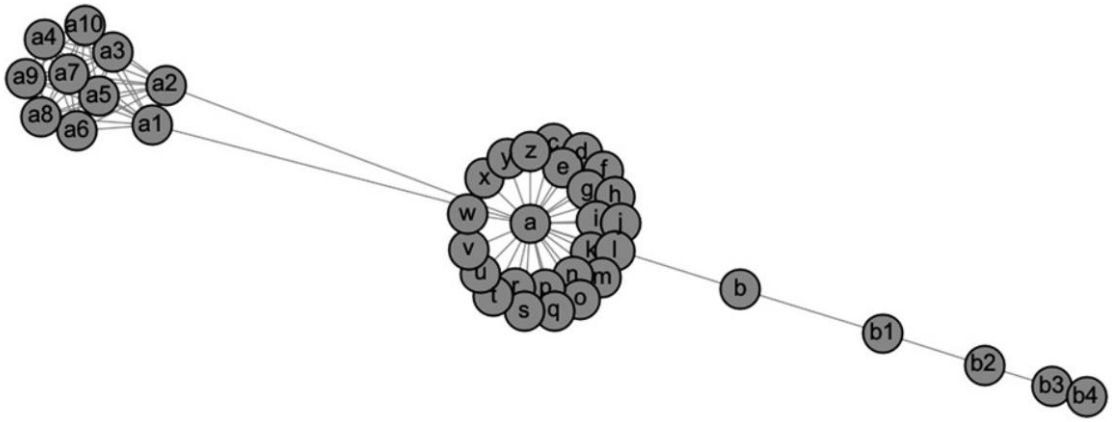

FIGURE 4.7

A layout of the sample network using the Force Atlas algorithm.

Yifan Hu layout

Many algorithms lay out graphs in this matter. Figure 4.3 uses one called Yifan Hu. Figure 4.7 uses a variant called Force Atlas. While there are differences between Figures 4.3 and 4.7, the similarities in clustering and separate nodes are clear.

Harel-Koren fast multiscale layout

The Harel-Koren fast multiscale algorithm (Harel and Koren, 2000), available in NodeXL, is designed to quickly lay out large, complex graphs. It is based on force-directed layout algorithms but uses optimizations in the underlying code to make the algorithm computationally efficient. For large graphs with thousands of nodes, generating a layout with many force-directed algorithms can take a very long time. With Harel-Koren, it often can be achieved in a few seconds, making it an ideal choice for large networks.

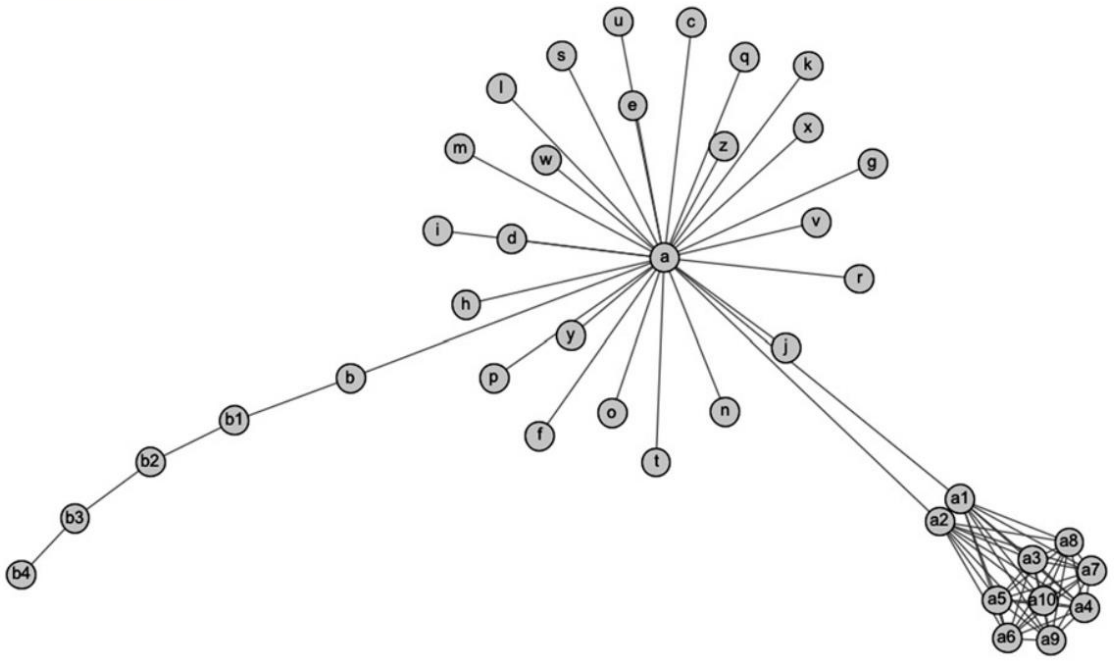

FIGURE 4.8

The graph laid out with the Harel-Koren Fast multiscale algorithm.

Other layouts

Most graphs will be presented using a force-directed layout algorithm. However, there are some more sophisticated layouts designed to convey additional information through layout. Figure 4.9 shows a layout available in the graphing program NodeXL. Here, nodes are clustered, grouped into boxes, and then links are added within and between boxes.

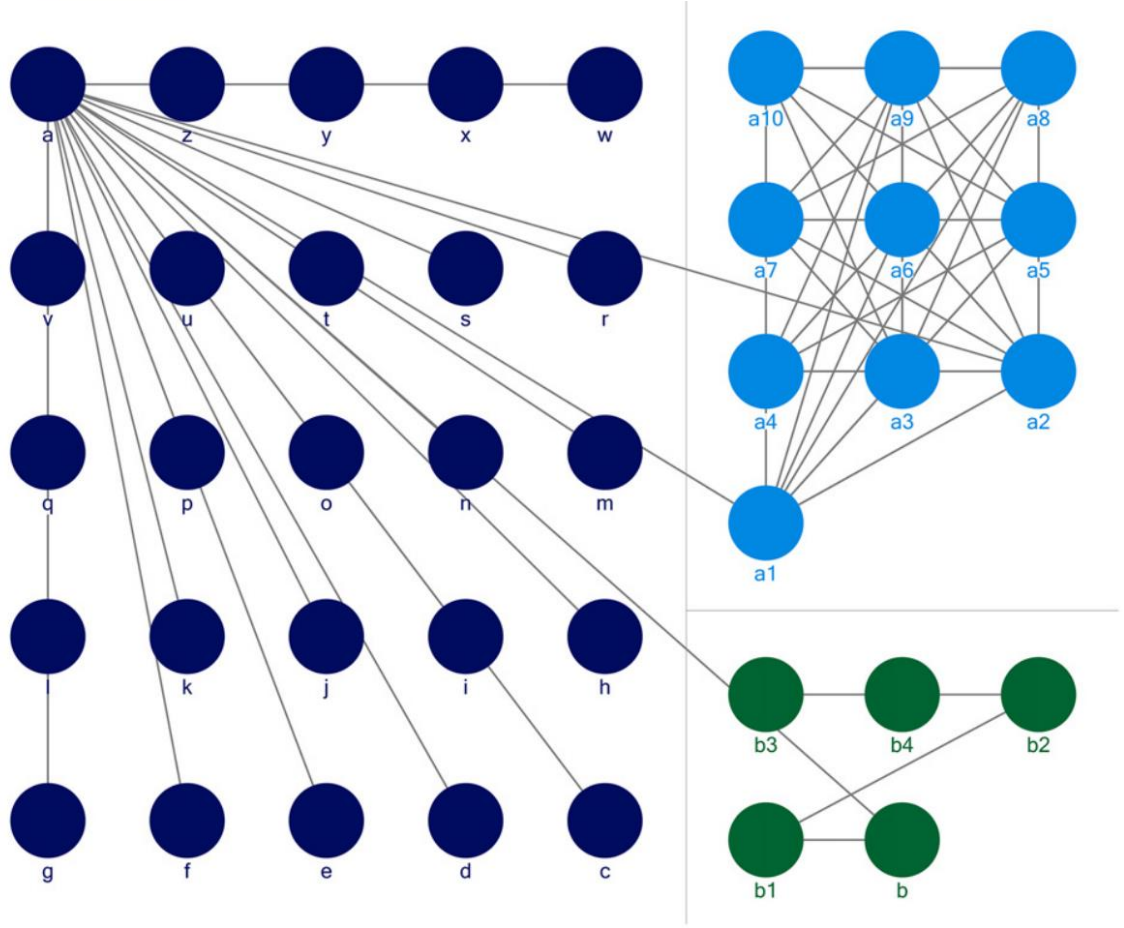

FIGURE 4.9

A layout that groups clusters into boxes, sized by the size of the cluster, and shows links between boxes.

Network visualization is an active area of research, so that new and creative mechanisms for visualization are constantly being developed. The examples above show the most commonly used and core methods of visualization, but network analysis tools will likely have additional options designed to support new and interesting types of analysis.

Visualizing network features

The layout algorithms discussed in the previous section dictate the placement of nodes and edges. Other network features, like edge weights, node properties, labels, and clusters, can also be visualized. Like the layout algorithms, there are many options to do this. This section will present some of the most common ways this is done.

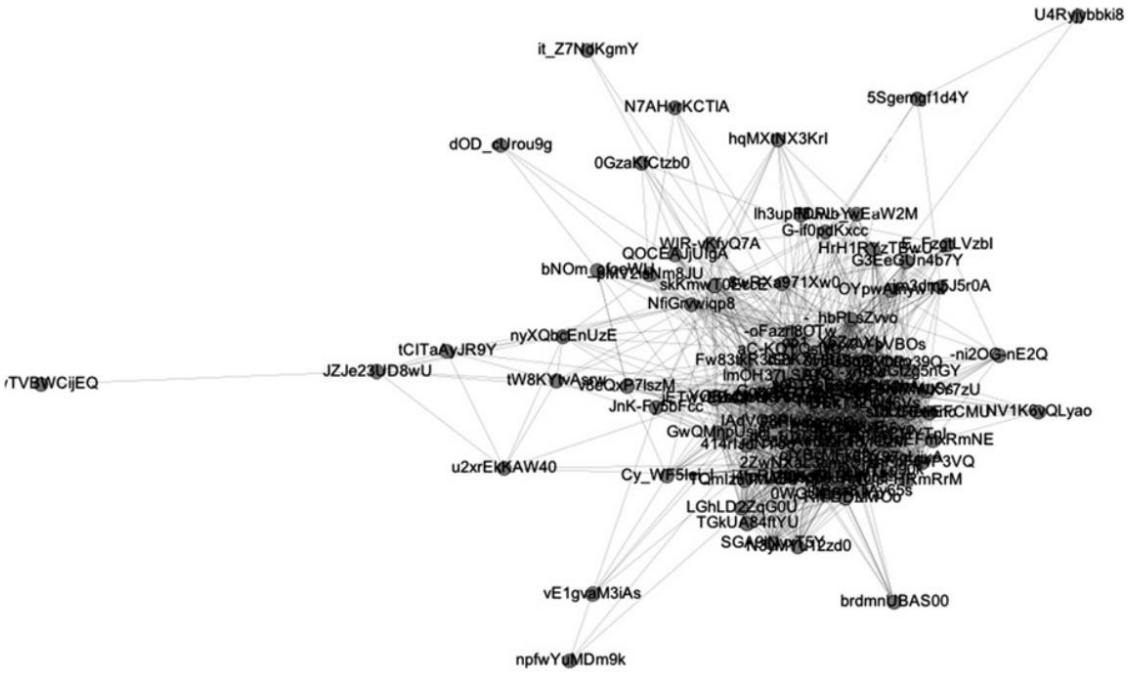

FIGURE 4.10

A network of YouTube videos with the node labels shown.

Labels

Labels are some of the more difficult attributes to show in a network, both on nodes and on edges. The example graphs in the previous section all have node labels, but the graphs are small and the labels are short. Figure 4.10 shows a network with only 92 nodes, which is still relatively small. The nodes represent YouTube videos, and the edges indicate that they were tagged with at least one similar term. The node labels are the YouTube identifiers for each video. Even in this small graph, the image becomes very cluttered with all the labels shown.

Similar problems happen with edge labels. Whether shown on top of the edge with straight alignment or angled along the edge, the graph tends to become cluttered and difficult to read. Some techniques can improve on this a bit, either by putting boxes around the text, by only showing a few labels of interest, or by relying on interactive interfaces that only show labels on demand. The latter allow the user to move the mouse over a node or edge and see the label or other data on demand. That facilitates exploration of the graph without the clutter. Still, there are no solutions to totally eliminate this problem when producing fixed visualization images, so often labels are left off.

Size, shape, and color

Fortunately, showing other attributes of nodes and edges in graphs can be easier. Categorical or quantitative attributes are particularly easy to show by adjustments in size, shape, or color. Return to the example graph used in Figures 4.3-4.8. There are many statistics about the nodes in that network: degree, centrality, and so on. These can be encoded using color, size, or both. Figure 4.11 shows color encoding of node degree. Darker colors indicate nodes with

higher degrees, and not surprisingly, node a is the darkest. For clarity, the node labels have been left off this graph.

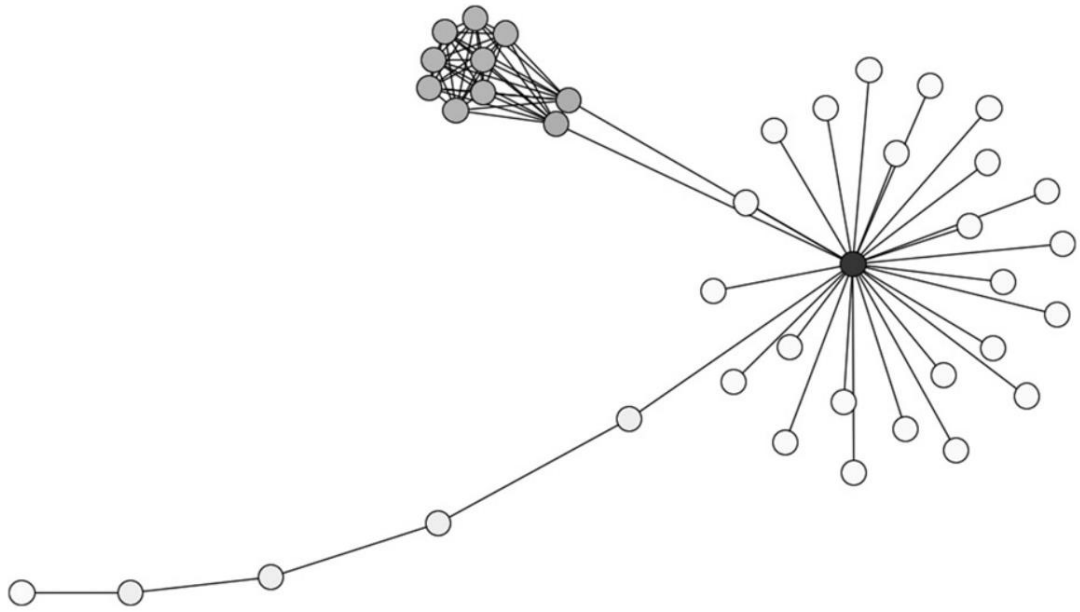

FIGURE 4.11

Color-coding nodes according to their degree, with higher degree shown by darker nodes.

Node color could also be used to indicate other attributes of a node. For example, in a visualization of a person's email network, node color could indicate if each person is a friend, family member, classmate, co-worker, and so forth.

Keeping color as an indicator of degree, node size can be used to indicate other attributes. For example, clustering coefficient is interesting here, since there is a tight cluster where all the nodes are connected, while in the rest of the graph, the clustering coefficient is very low for each node. Figure 4.12 shows a graph that uses color for degree and size for clustering coefficient.

Edges can also be treated with color or thickness to indicate their attributes. For example, different types of relationships could each be coded in a different color. Edge weights are also commonly visualized. These could indicate the strength of a relationship, the frequency of communication, or other factors. Figure 4.13 shows the same example network with weights added to the edges. These are visualized by adjusting the width of the edge. Wider edges indicate stronger relationships.

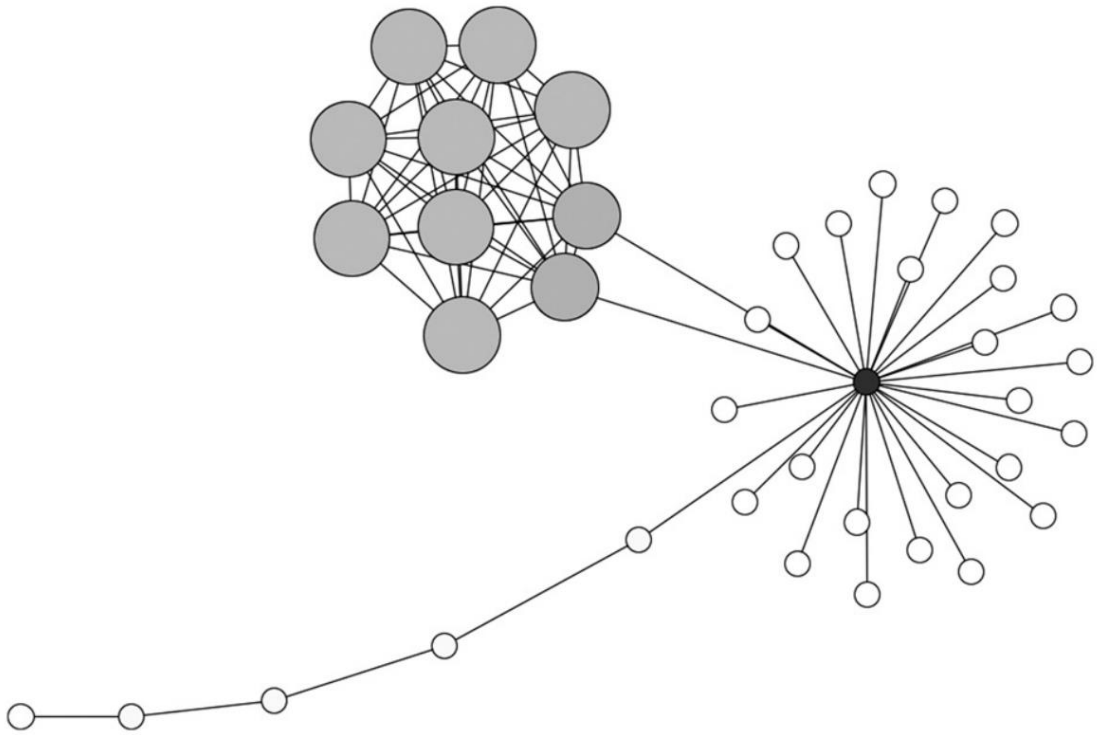

FIGURE 4.12

A graph indicating clustering coefficient with node size and degree with node color.

Larger graph properties

Larger graph properties can also be encoded in visualizations. For example, clusters are sometimes apparent on their own (like the group to the upper right in Figure 4.11), but visual properties to indicate them will often clarify a visualization further. Figure 4.13 shows a new graph that has two main clusters. This graph is a network of YouTube videos, where nodes represent videos and edges connected videos that share a common tag. All of these videos were tagged with the word “cubs”; this example will be discussed more in Chapter 7. Even without the color coding, the two groups would be relatively easy to see. But using a community detection algorithm that groups nodes into clusters, and then color coding by those clusters, makes it even more apparent. This is shown in Figure 4.14.

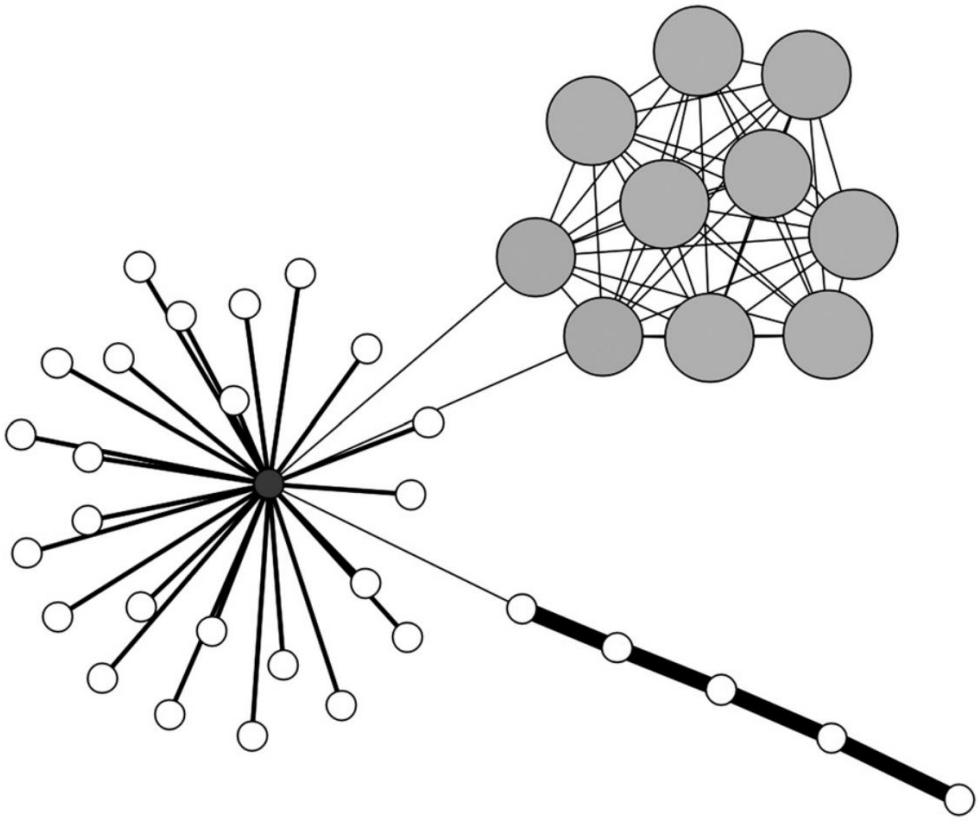

FIGURE 4.13

The sample network with edge width indicating the weight on each edge. Note that the central node has medium-strength relationships with most neighbors, but weak ones to the cluster in the upper right and the chain in the lower right. The chain of nodes in the lower right have high weights on the edges connecting them.

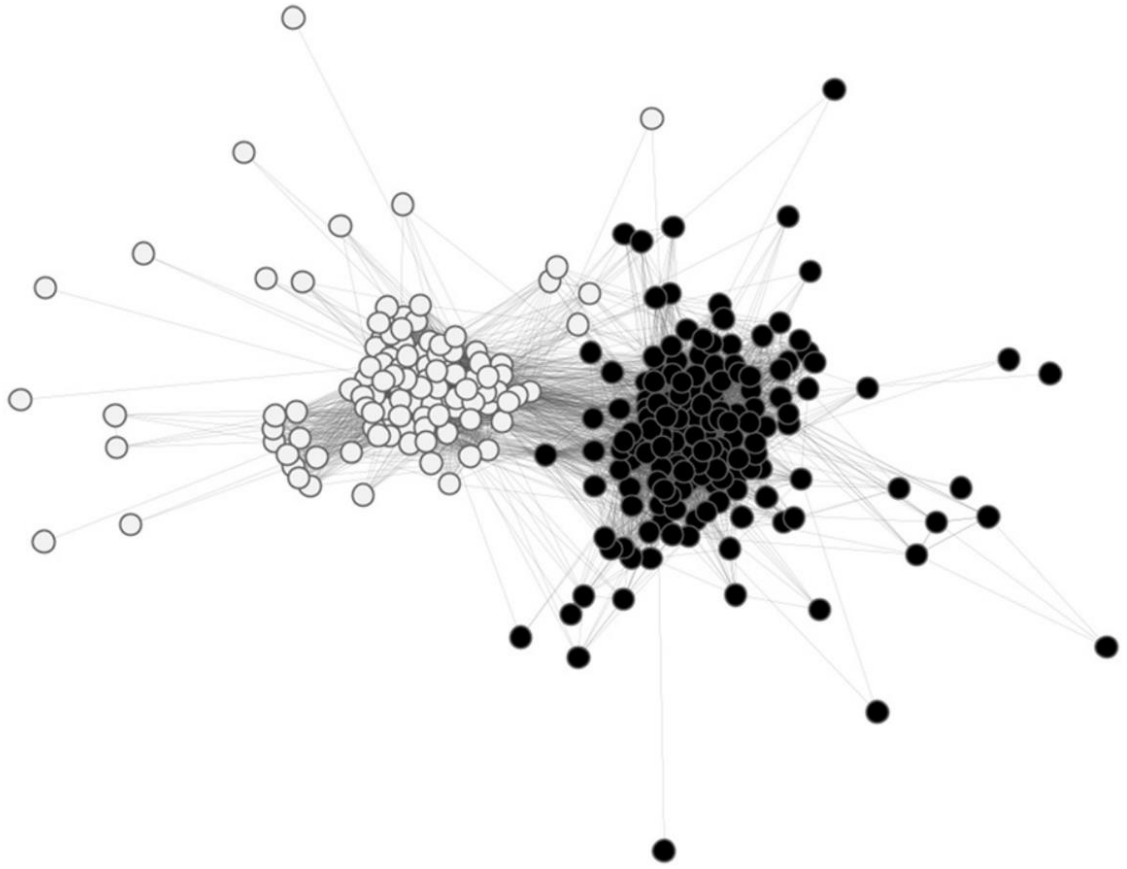

FIGURE 4.14

A network of YouTube videos where color indicates the community or cluster to which each node belongs.

Scale issues

The example networks shown so far have been relatively small a few hundred nodes and a few thousand edges. Visualization is very useful for analyzing networks of this size or smaller. When networks become much larger, the quality of the visualization diminishes.

Figure 4.15 shows a network from a peer-to-peer file sharing network. Nodes represent hosts (computers participating in the network), and edges represent connections between them (usually one computer downloading a file from another). There are close to 11,000 nodes in this network with roughly 40,000 edges. Even with a very low density (<0.001), there are still too many nodes and edges to see much of anything.

Depending on the structure of the network, it is sometimes possible to get useful visualizations with up to around 10,000 nodes; however, networks under 1,000 nodes are typically safest.

Density

Density can also be a problem for visualization, even if the number of nodes is small. Figure 4.16 shows a network of members of the U.S. Senate. There are only 100 nodes but over 4,100 edges. The edges indicate that the senators have voted the same way in at least one bill.

The edges have a weight, indicating the percentage of bills on which the two senators have voted in the same way. Figure 4.13 has the edges filtered so that only those with a weight of 40% or more are visible. However, as this network shows, there are no interesting patterns visible with the threshold of 40%; the network is simply too dense.

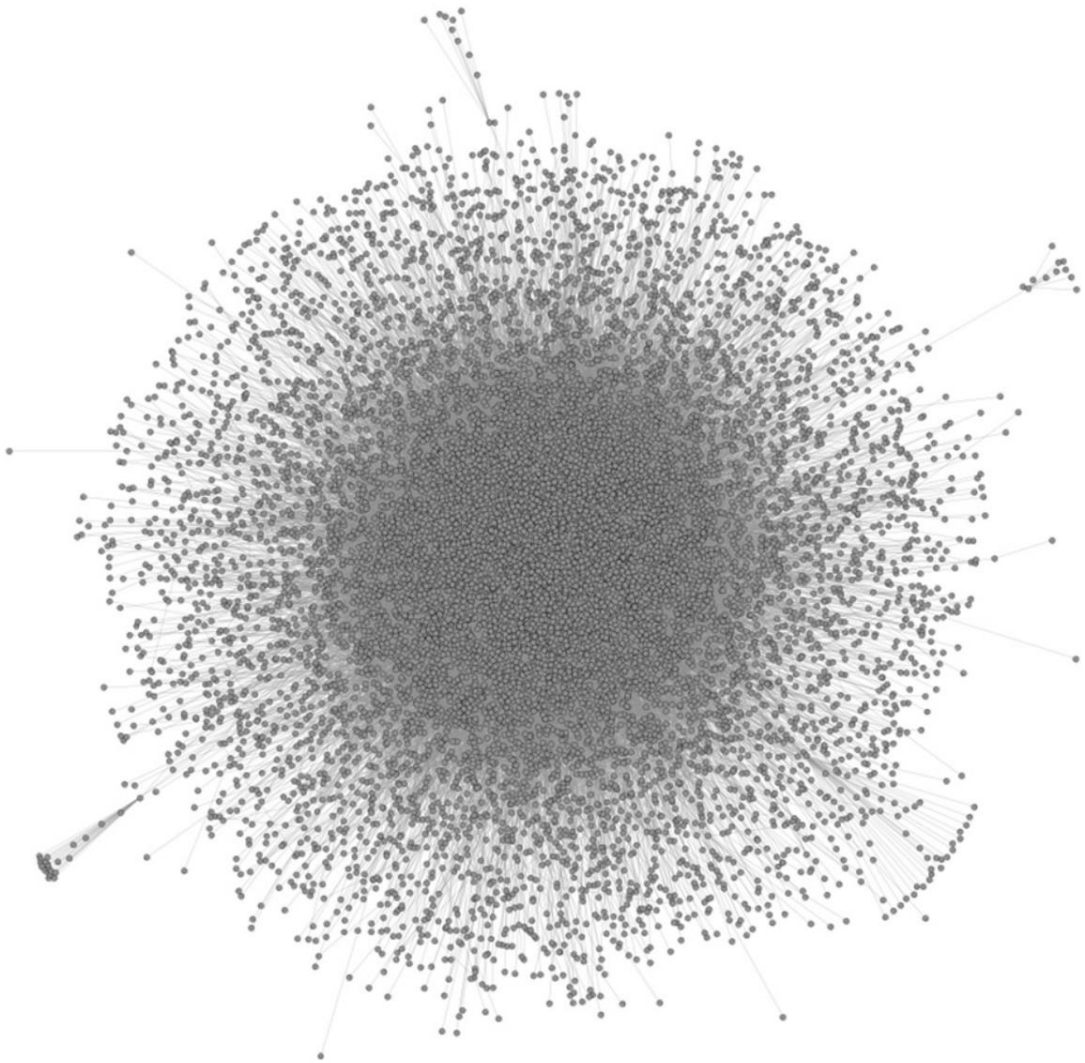

FIGURE 4.15

A network with 11,000 nodes and 40,000 edges.

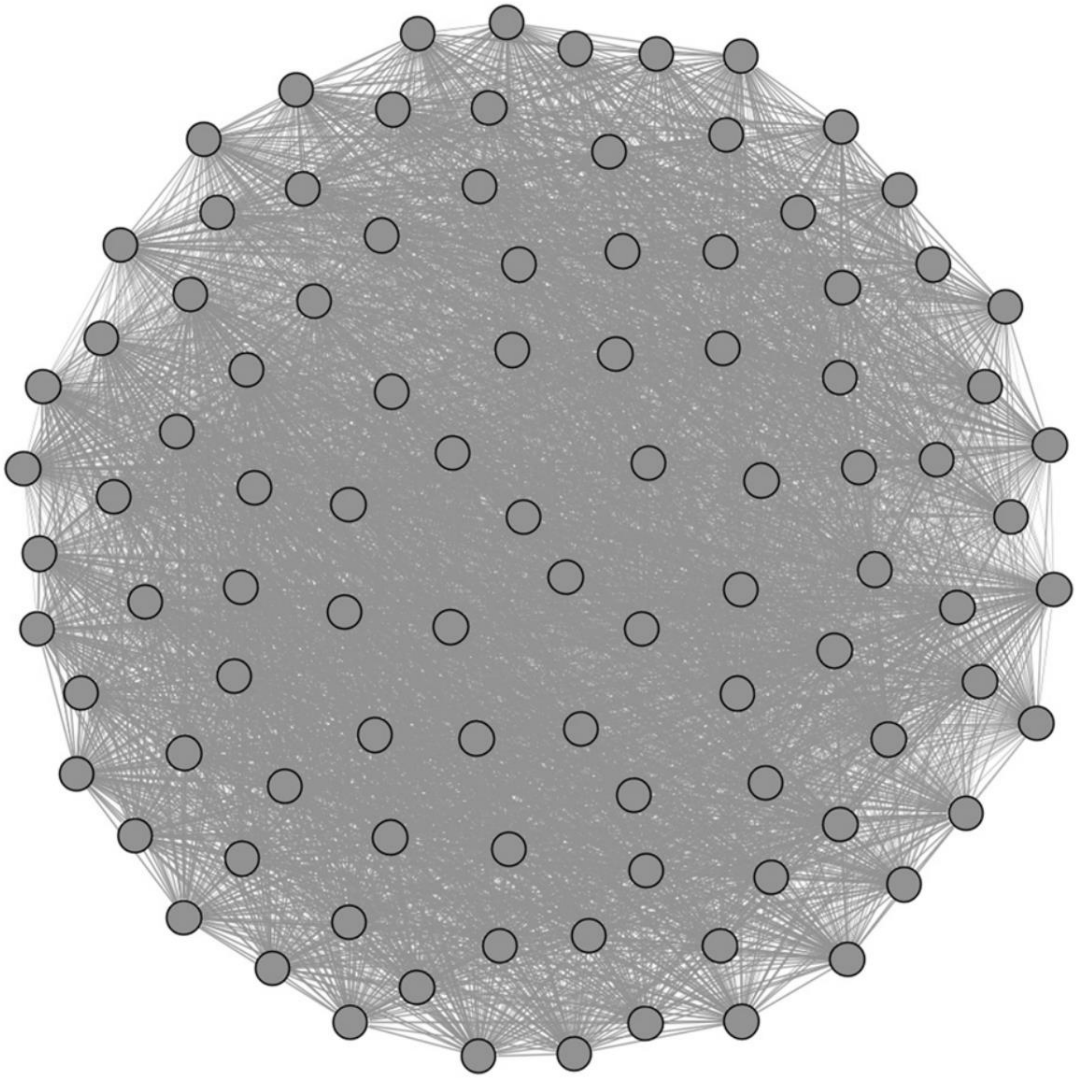

FIGURE 4.16

A network of senators (nodes) with edges connecting senators who have voted the same way at least 40% of the time. The network is very dense, so it is not possible to see any interesting patterns.

Tie Strength:

Social relationships are complicated. The type of relationship people have will draw on many things like their history and similarity, each person's personal background and preferences, environmental factors, and more. Relationships are also multifaceted, and many relationship types can be used in social network analysis. One of the most useful is the idea of tie strength.

Tie strength is a measure of the strength of a relationship between people. The concept was introduced by Mark Granovetter in 1973. He asserted that "the strength of a tie is a ... combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie" (Granovetter, 1973).

While there is a range of tie strength, Granovetter defined two main types: *strong ties* and *weak ties*. Strong ties are rare and are usually family members or very close friends. These are usually people a person sees frequently, with whom one shares personal details of one's life, and for whom the person will do and expect favors. Weak ties are much more common and include acquaintances and more casual friendships. Of course, there is a spectrum of tie strength, and any relationship may fall along the scale from weak to strong.

People with whom someone has no meaningful relationship the familiar stranger one passes on the street and nods to, or a vendor that a businessperson may contact are not considered in the spectrum of strong or weak since there is not much of any relationship present. These are sometimes called *absent ties* and would not appear as edges in the social network.

Tie strength is a very important factor to consider in social network analysis. Consider the flow of information through a network. Weak ties often connect to diverse groups of people with different perspectives. These ties allow information to move throughout the network. Strong ties are more trusted, and their information is more likely to be reliable. The same features apply when considering the spread of other things through a network, like a disease. Someone is more likely to catch a cold from a weak tie (because there are many of them, and they will carry germs from many different groups of people). But because of the high level of close contact, it will likely spread quickly to one's strongest connections.

That is not to say that tie strength is the only factor influencing trust, reliability, and closeness in social networks. Weak ties may provide highly trusted information; for example, a physician may be more trusted about medical information than someone's family members. In this example, the authority of the physician

outweighs tie strength. Throughout the chapter, we will discuss factors that influence tie strength and its role, but there will be exceptions to every example. Thus, tie strength cannot be treated as the only factor influencing relationships, and observing people's interactions cannot predict it perfectly, but the guidelines presented here are useful for considering this important relationship measure and its role in networks.

Measuring tie strength

To analyze tie strength in social network analysis, the network must include relationship information. In small networks, especially if data is hand-collected, it may be feasible to ask each person to rate the strength of their tie to each person. By necessity, larger networks require a mechanism for measuring tie strength. There is no single factor that defines a strong or weak tie, but a number of predictors can be combined to estimate the strength of a relationship.

In his original paper, "The Strength of Weak Ties," Granovetter offers four intuitive factors that may contribute to tie strength. As stated above, he writes, "the strength of a tie is a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."

Time can include the amount of time people spend with each other, the duration of their relationship (i.e., how long they have known each other), and how frequently they see one another.
Emotional intensity is indicated by the closeness of a relationship; close friends or family members are likely to be strong ties, while more casual friends and acquaintances would be weaker ties.
Intimacy, or mutual confiding, relates to people sharing secrets or intimate personal details with one another. The more of this information they exchange, the closer their relationship is likely to be.
Reciprocal Services are favors that people do for one another. They may be personal (e.g., pet sitting or picking up someone's dry cleaning), financial (e.g., loaning money), professional (e.g., putting people in contact with one another), or otherwise.

Since originally proposed in 1973, researchers have investigated what other factors might also play a role in tie strength. There are several of these factors, but three are more widely accepted as important.

Structural features relate to the social network of the two people in question. Those who have many mutual friends are likely to have stronger ties.
Social Distance measures how different people's social situations are. This includes factors like age difference, race, education, and socioeconomic status. People with strong ties tend to have similar social attributes.

Emotional Support describes the communication between people that validates their emotions, shows understanding of their problems, and tries to alleviate stress.

These seven factors are not equally important in determining tie strength (although there is not total agreement about their relative importance). For example, studies have consistently shown that measures of a relationship's closeness, often captured through emotional intensity or intimacy, are among the strongest indicators of strength (Marsden, 1984).

These factors are not independent. For example, people who have a very intimate relationship will often spend a lot of time together. People of different ages and positions in life, or those who have a large social distance are also less likely to have as many mutual friends as people with similar social positions. Thus, when measuring behavior or interactions, a single measurement may describe more than one of these factors.

Additionally, it does not always follow that having many of these factors indicates a strong tie. For example, roommates may have many friends in common, be in socially similar situations (and therefore have a low social distance), spend a lot of time together, and even do favors for one another, yet still maintain a distant and impersonal relationship.

A natural question to follow is how these factors are measured. Intimacy, for example, is difficult to quantify, and depending on the context of a relationship, its meaning may vary. Indeed, there is no single correct answer for how to measure any of these relationship features. If measuring them is important, it will depend on the context, the information available, and likely many other factors.

An interesting example of one way this measurement has been done is presented in Gilbert and Karahalios's work on computing tie strength in social media (2009). In their study, subjects answered a series of questions about their relationship with friends on Facebook, and information was collected from both users' profiles and their interactions. This Facebook data was used to create a set of attributes designed to reflect each of the seven aspects of tie strength mentioned above.

Here are just a few examples of the over 70 variables they used to measure tie strength in their study.

Intimacy

• Number of days since their last communication

• Number of friends in common

• Number of "intimate" words in their communications, as determined by software that automatically analyzes text

Intensity

• Number of words exchanged on one another's walls

• Depth of email threads in their inboxes (i.e., how many messages were sent back and forth in a conversation)

Reciprocal services

• Number of links shared on one another's wall

• Applications the users had in common (presumably because they could be working together within the application context)

Social distance

- Age difference
- Difference in the number of educational degrees
- Difference in the number of occupations

Tie strength and network structure

Strong ties have unique properties within a social network. They are not randomly scattered throughout the network, but rather tend to appear in clusters. As an exercise, think of five to seven of your strongest relationships. Write these in a circle, and draw connections between the people who have relationships with one another. Use thicker lines for the strong ties between these people. Very often, there will be many strong ties among the people with whom you share strong ties.

This illustrates the tendency of strong ties to appear in clusters. Each person will have many more weak ties connecting them to people outside this small group, but a person's strong ties tend to have strong ties to one another.

This pattern of strong ties being densely connected leads to another structural concept called the *forbidden triad* (see Figure 5.2). Imagine three people: Alice, Bob, and Chuck. Alice and Bob have a strong tie, and Alice and Chuck also have a strong tie. What does that tell us about the relationship with Bob and Chuck? While we cannot draw any absolute conclusions, it is likely that some sort of tie exists between Bob and Chuck, either strong or weak. When that tie does not exist, it is known as the Forbidden Triad.

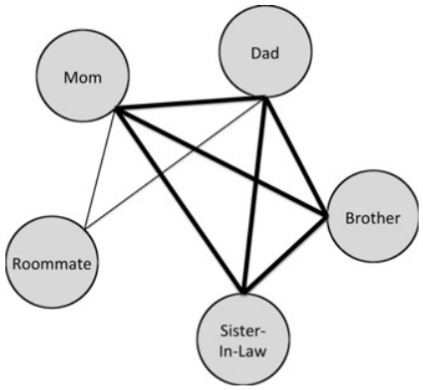

FIGURE 5.1

Sample Exercise. Note that there are strong ties connecting four of the five people listed, and two more weak ties. Only two ties are absent, between the roommate and brother, and roommate and sister-in-law. This is a very densely connected network with many strong ties.

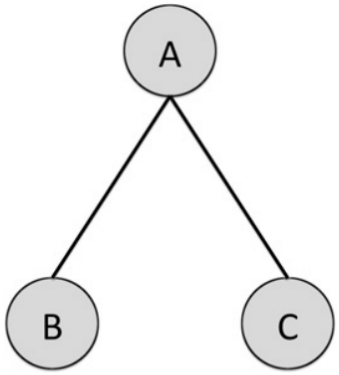

FIGURE 5.2

The Forbidden Triad.

Granovetter named this triad "forbidden" because of the unlikelihood that no connection between Bob and Chuck exists. It is an exaggeration to say this never occurs, but studies have shown that it occurs less frequently than one would expect if tie strength were randomly distributed between people in the network. One can also think of this structure as representing something actually forbidden, like a person (A) who is married (to B) and is also having an affair (with C).

From this, a second principle relating network structure and tie strength arises. Consider Figure 5.3.
In this network, P and F have a strong tie connecting them. This is the only edge that connects F's cluster of nodes to P's cluster of nodes, so it is a bridge. Recall that a bridge is an edge that is the only connection between two groups of nodes.

Nodes P and F have other strong ties as well. Indeed, in almost all social networks, nodes have more than one strong tie. In this network, F has a strong tie to O, and P has strong ties to H and N. We can form three triads with P, F, and another node where there are two strong ties: PFO, PFH, and PFN. In these cases, if there is no third connection, we are left with a forbidden triad. For example, we expect that there should be an edge between P and O since strong ties connect F

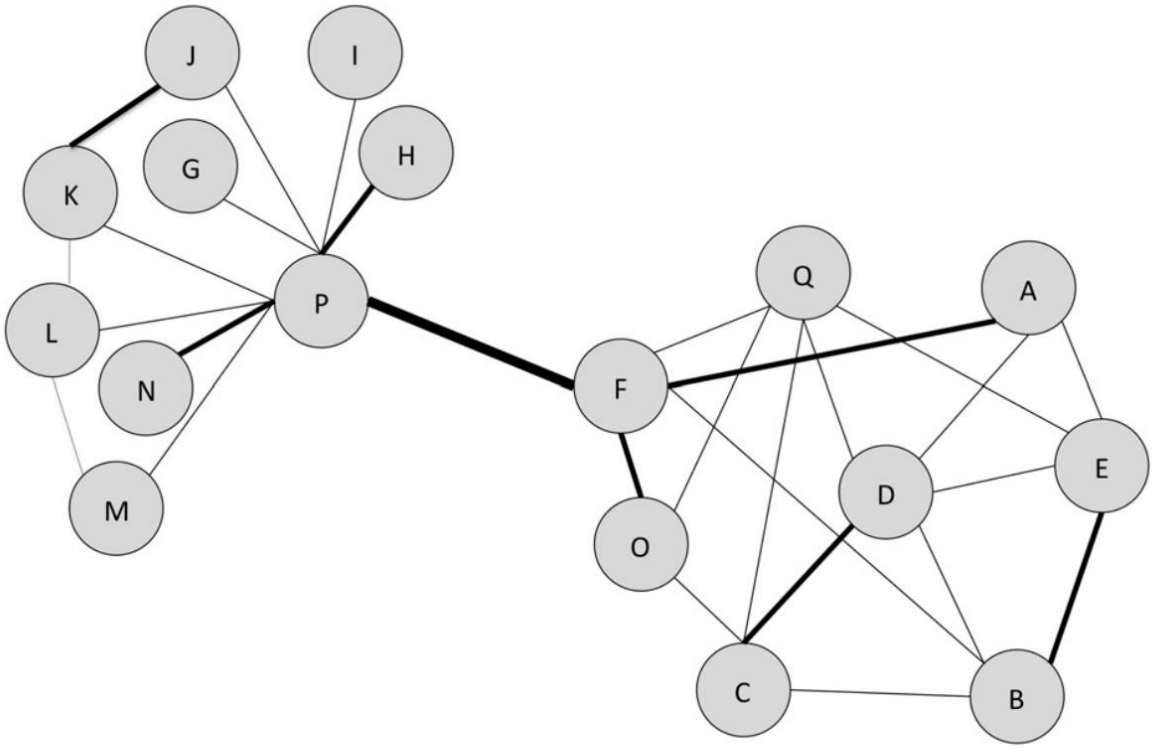

FIGURE 5.3

The edge between P and F is a bridge that connects the two clusters of nodes. It is a strong tie, and thus we would expect connections between some of the triads with two strong ties (e.g. PFO, PFH, PFN). It is very unlikely that no tie third tie would exist in any of those triads, and thus it is unlikely that a strong tie would be a bridge.

to both P and O. If such an edge were to exist, even as a weak tie, then the edge between P and F would no longer be a bridge; the new edge (e.g., an edge between P and O) would be another path connecting the two clusters. Thus, since several forbidden triads are unexpected, it is very unlikely that a strong tie will ever be a bridge; another edge is likely in one of these triads, and that will add another connection between the clusters. Granovetter described this in his work with the principle that no strong tie is a bridge; while strong ties may be bridges, it is unlikely given what we know about the distribution of edges. It is also unlikely that, over time, a strong tie would remain a bridge. Weak ties would be likely to form and connect nodes to remove the strong tie's bridge status.

Trust

This is a fundamental part of a trust relationship. The person being trusted is expected to do the "right" thing. This usually means she will act with the other person's best interests in mind and/or take actions that benefit the other person.

The person making the decision about whether or not to trust someone is considering more than just her expectations about the other person's actions. She must also decide if she is willing to take some risk by putting her trust in the other person. That may be a small risk or a large one. Receiving and acting on a poor movie recommendation may only waste a few hours of time, but making a large loan that is not repaid can have major effects. So can asking for a recommendation letter from someone who will not write a good one.

All of these ideas can be condensed down into several important factors. First, the person doing the trusting must make herself vulnerable and take some risk by trusting the other person. Second, she takes that risk because she believes the other person will act well or behave in a way that will benefit her. Vulnerability, risk, and positive expectations of the other person are the core components of the trust relationship.

Thus, as a definition of trust, we can say the following: A person trusts another if she is willing to take a risk based on her expectation that the trusted person's actions will lead to a positive outcome.

Development of trust

Trust is formed between people in a wide variety of ways. In a common scenario for building trust, one person develops trust in another over time through a series of interactions that help the person build up a belief in the reliability and good intentions of the other, eventually to the point where she is willing to take a risk and act on the building trust. A series of risks that are rewarded lead to more trust. However, this does not always happen and is often not possible. Consider meeting a physician for the first time. Someone may trust that doctor with his health, but it is based on factors such as background, qualifications, references from other people, and personal compatibility rather than on a series of successful interactions over time. Someone may also have to immediately develop trust in another person, such as a victim trusting a rescuer in an emergency.

McKnight et al. (1998) document four major components of the way people consider and build trust in others:

1. **Calculation-based trust**: This is a rational decision about whether to trust someone, and where the costs and benefits of trusting are factored in.
2. **Personal-based trust**: This reflects a person's propensity to trust, developed over the course of their life.
3. **Cognition-based trust**: This describes the instant rapport and trust that can develop between people who share similar backgrounds, beliefs, and values. It often is based on first impressions.

4. **Institution-based trust**: This addresses how trust may form in the presence of guarantees and protections offered by an institution.

These factors can be applied in a wide range of contexts, including the study of trust between people and from people to organizations or communities.

Asymmetry

For two people involved in a relationship, trust is not necessarily identical in both directions. Because individuals have different experiences, psychological backgrounds, and histories, two people may trust each other at different levels. For example, parents and children clearly trust one another differently. Children must have almost absolute trust in their parents, while the parents may have almost no trust in their children, particularly when they are very young.

Context and time

Except for a few of these very asymmetric relationships, like that between parents and young children, trust is rarely all-encompassing. Rather, a person will tend to trust someone else about a set of things, but not about everything. For example, someone may trust her friend to recommend a movie but not to repair her car. She may trust her boss to edit a document but not to perform surgery on her. When people build trust in one another, and when they rely on it, it is usually connected to those contexts. However, trust may sometimes transfer from one context to another. A person may build trust in a co-worker that is entirely in the work context, but later trust that person to recommend a plumber, even if they have never had a discussion about plumbing or household repair.

Measuring trust

Measuring trust is important but difficult. People perceive trust differently, and trust is also difficult to quantify or explain. When studying how to measure trust, it has generally been broken down into two parts: a person's propensity to trust, and one person's decision about the other person.

1. **Propensity to trust**
   - Refers to a **person's inclination to trust others.**
   - Investment Game is a common way to measure a person's propensity to trust and the trustworthiness of others

2.Trust in others

- An individual's amount of trust in others varies based on their propensity to trust and their beliefs
- about the specific person they are deciding to trust.
- Decisions about trust in others deal with the risks a person is willing to take with a specific person.
  - 1. Trust with material possessions
  - 2. Belief about reliability
  - 3. Trust with secrets
  - 4. Trust regarding physical safety

- **Trust in social media** is **harder to judge** compared to in-person interactions as the information available about a person is limited and identity can be easily forged.
- **Before social media**, users were mostly concerned with trusting websites, especially **e-commerce sites,** and retailers tried to overcome this by creating **privacy policies,** assuring transaction safety and security, and building professional and secure websites.
- With the rise of social media and user interaction, **trust issues shifted to include both websites and other users.**
- **Reputation systems,** such as the one **used by eBay**, were introduced to help users make decisions about **trust based on feedback from other users**.
- **Social media users often share a lot of personal information**, some of which is private, and must trust both the website and other users to treat their information with respect.
- **Complex issues** regarding trust in social media arise due to the constantly changing factors of **what users want to share, with whom they want to share it, and what control they have over those decisions.**
- Reputation systems are not always present in social media applications and understanding trust in social media is a complex and constantly evolving topic.
- The problem of inferring trust arises in social networks, where trust between two people can be considered a **weight on the edge connecting them.**
- **The determination of trust** is influenced by **an individual's propensity to trust and their assessment of the other person's trustworthiness.**
- Surveys could be used to generate a trust rating, but most people will not take the time to fill them out.

SOCIAL MEDIA NETWORK ANALYTICS

Networks are the building blocks of social media and can carry useful business insights. Social media network analytics thus deals with constructing, analyzing, and understanding social media networks. Social network analytics can be used for variety for purposes. It can be employed to identify influential nodes (e.g., people and organizations) or their position in the network, or to understand the overall structure of a network. An organization may mine their Twitter or Facebook follower networks to identify influential network leaders and empower them. People occupying central positions in social media networks are of great value to social marketers, as they have the ability to propagate information to vast numbers of people and are considered opinion leaders.

The purpose of network analysis is to:

✓Understand overall network structure; for example, number of nodes, number of links, density, clustering coefficient, and diameter.

✓ Find influential nodes and their rankings; for example, degree, betweenness, and closeness centralities.

✓ Find important links and their rankings; for example, weight, betweenness, and centrality.

✓ Find cohesive subgroups; for example, pinpointing communities within a network.

✓ Investigate multiplexity; for example, analyzing comparisons between different link types, such as friends vs. enemies.

COMMON NETWORK TERMS

NETWORK At a very basic level, a network is a group of nodes that are connected with links (Wasserman and Faust 1994). Nodes (also known as vertices) can represent anything, including individuals, organizations, countries, computers, websites, or any other entities. Links (also known as ties, edges, or arcs) represent the relationship among the nodes in a network. Networks can also exist among animals.

SOCIAL NETWORKS A social network is a group of nodes and links formed by social entities where nodes can represent social entities such as people and organizations. Links represent their relationships, such as friendship and trade relations. Social networks can exist both in the real and online worlds. A network among classmates is an example of real world social network. And a Twitter followfollowing network is an example of an online social media network. In a Twitter follow-following network, nodes are the Twitter users, and links among the nodes represents the follow-following relationship (i.e., who is following whom) among the users.

SOCIAL NETWORK SITE A social network site is a special-purpose software (or social media tool) designed to facilitate the creation and maintenance of social relations. Facebook, Google+, and LinkedIn are examples of social network sites.

SOCIAL NETWORKING The act of forming, expanding, and maintaining social relations is called social networking. Using social network sites, users can, for example, form, expand, and maintain online social ties with family, friends, colleagues, and sometimes strangers.

SOCIAL NETWORK ANALYSIS Social network analysis is the science of studying and understanding social networks (Hanneman and Riddle 2005) and social networking. It is a well established field with roots in a variety of disciplines including Graph Theory, Sociology, Information Science, and Communication Science.

COMMON SOCIAL MEDIA NETWORK TYPES

The following are some everyday types of social media networks that we come across and that can be subject to network analytics.

FRIENDSHIP NETWORKS The most common type of social media networks are the friendship networks, such as Facebook, Google+, and Cyword. Friendship networks let people maintain social ties and share content with people they closely associate with, such as

family and friends. Nodes in these networks are people, and links are social relationships (e.g., friendship, family, and activities).

FOLLOW-FOLLOWING NETWORKS In the follow-following network, users follow (or keep track of) other users of interested. Twitter is a good example of follow-following network where users follow influential people, brands, and organizations. Nodes in these networks are, for example, people, brands, and organizations, and links represents follow-following relations (e.g., who is following whom).
Below are two common Twitter terminologies.
Following—Following are the people who you follow on Twitter. Following someone on Twitter means:

- You are subscribing to their tweets as a follower.
- Their updates will appear in your Home tab.
- That person is able to send you direct messages.

Followers—Followers are people who follow you on Twitter. If someone follows you, it means that:

- They will show up in your followers list.
- They will see your tweets in their home timeline whenever they log in to Twitter.
- You can send them direct messages.

FAN NETWORK A fan network is formed by social media fans or supporters of someone or something, such as a product, service, person, brand, business, or other entity. The network formed by the social media users subscribed to your Facebook fan page is an example of a fan network. Nodes in these networks are fans, and links represent colikes, cocomments, and coshares.

GROUP NETWORK Group networks are formed by people who share common interests and agendas. Most social media platforms allow the creation of groups where member can post, comment, and manage in-group activities. Examples of social media groups are Twitter professional groups, Yahoo Groups, and Facebook groups. Nodes in these networks are group members, and links can represent cocommenting, coliking, and coshares.

PROFESSIONAL NETWORKS LinkedIn is a good example of professional networks where people manage their professional identify by creating a profile that lists their achievements, education, work history, and interests. Nodes in these networks are, for example, people, brands, and organizations, and links are professional relations (such as coworker, employee, or collaborator). An important feature of professional networks is the endorsement feature, where people who know you can endorse your skills and qualification.

CONTENT NETWORKS Content networks are formed by the content posted by social media users. A network among YouTube videos is an example of a content network. In such

a network, nodes are social media content (such as videos, tags, and photos) and links can represent, for example, similarity (content belonging to the same categories that can be linked together).

DATING NETWORKS Dating networks (such as match.com and Tender) are focused on matching and arranging a dating partner based on personal information (such as age, gender, and location) provided by a user. Nodes in these networks are people, and links represent social relations (such as romantic relation).

COAUTHORSHIP NETWORKS Coauthorship networks are two or more people working together to collaborate on a project. Wikipedia (an online encyclopedia) is a good example of a social media-based coauthorship network created by millions of authors from around the world(Biuk-Aghai 2006). A more explicit example of the coauthorship network is the ResearchGate platform: a social networking site for researchers to share articles, ask and respond to questions, and find collaborators. In these networks, nodes are, for example, researchers, and links represent the coauthorship relationship.

COCOMMENTER NETWORKS Cocommenter networks are formed when two or more people comment on social media content (e.g., a Facebook status update, blog post, or YouTube video). A cocommenter network can, for example, be constructed from the comments posted by users in response to a video posted over YouTube or a Facebook fan page. In these networks, nodes represent users, and link represents the cocommenting relationship.

COLIKE In a similar way, colike networks are formed when two or more people like the same social media content. Using NodeXL (a social network analysis tool), one can construct a network based on colikes (two or more people liking a similar content) over Facebook fan page. In such network, nodes will be Facebook users/fans and links will be the colikes relationship.

COOCCURRENCE NETWORK Cooccurrence networks are formed when two more entities (e.g., keywords, people, ideas, and brands) cooccur over social media outlets. For example, one can construct a cooccurrence network of brand names (or people) to investigate how often certain brands (or people) cooccur over social media outlets. In such networks, nodes will be the brand names and the links will represent the cooccurrence relationships among the brands.

GEO COEXISTENCE NETWORK Geo coexistence networks are formed when two more entities (e.g., people, devices, and addresses) coexist in a geographic location. In such a network nodes represents entities (e.g., people), and links among them represent coexistence.

HYPERLINK NETWORKS In simple words, hyperlink is a mechanism to move among electronic documents (such as websites). Hyperlinks can be referred to as being either in-links (i.e., hyperlinks originating in other websites (Björneborn and Ingwersen 2004), thus bringing traffic/users to your website) or out-links (i.e., links originating in your website and going out (Bjorneborn 2001), thus sending traffic to other websites). Hyperlink also forms networks. Typically, in these network nodes are website, and links represent referral relationships (in the form of in-links or out-links).

TYPES OF NETWORKS

From a technical point of view, the above-mentioned networks can be classified in a variety of ways, including 1) based on existence, 2) based on direction of links, 3) based on mode, and 4) based on weights.

BASED ON EXISTENCE

Based on the way the networks exist online or are constructed, they can be classified as 1) implicit networks or 2) explicit networks.

1. Implicit Networks Implicit networks do not exit by default (or are hidden) and need to be intentionally constructed with the help of dedicated tools and techniques. Examples of such networks include keyword cooccurrence networks, cocitation networks, cocommenter networks, hyperlink networks, etc. Constructing and studying implicit networks can provide valuable information and insights.

2. Explicit Networks Explicit social media networks exist by default; in other words, they are explicitly designed for social media users to be part of. Most social media networks are explicit in nature. Examples of explicit social media networks include Facebook friendship network, Twitter follow-following networks, LinkedIn professional networks, YouTube subscribers' network, and bloggers' networks. In this chapter we will focus on explicit social media networks.

BASED ON DIRECTION Based on the directions of links among the nodes, the networks can be classified as 1) directed networks, and 2) undirected networks.

1.Directed Networks A network with directed links among nodes is called a directed network (Figure 6). Usually, a link with an arrow is drawn to show the direction of the relationship among the nodes. For example, the Twitter following-following network is a directed network where direction of the arrow shows who is following whom.

2.Undirected Network In undirected networks, the links among the nodes do not have any direction. A Facebook friendship network is an example of undirected network.

BASED ON MODE Based on the composition of nodes, networks can be classified as 1) one-mode network, 2) two-mode networks, and 3) multimode networks.

1.One-Mode Networks A one-mode network is formed among a single set of nodes of the same nature (Figure 6). A Facebook friendship network is an example of a one-mode network where nodes (people) form network ties (friendships).

2.Two-Mode Networks Two-mode networks (also known as bipartite networks) are networks with two sets of nodes of different classes (Latapy, Magnien et al. 2008). In these networks, network ties exist only between nodes belonging to different sets (Figure 6). For example, consider the two-mode network given in Figure 6, where one set of nodes (circles) could be social media users and other set of nodes (squares) could be participation in a series of events. Users are linked to the events they attended.

3.Multimode Network A multimode network is also possible where multiple heterogeneous nodes are connected together. It can be considered as an amalgam of one and two-mode networks.

BASED ON WEIGHTS Networks can also be classified based the weight assigned to the links among the nodes. Mainly there are two types of weighted networks: 1) weighted networks, and 2) unweighted networks.

1.Weighted Networks In weighted networks, the links among nodes bear certain weights to indicate the strength of association among the nodes. The link (relationship) between, for example, two Facebook friends (nodes) will be thicker if they communicate more frequently (Figure 6). Weighted networks can provide rich information, but are difficult to construct.

2.Unweighted Networks In unweighted networks, links among nodes does not bear weights. The links only indicate the existence of a relationship and cannot provide clues about the strength of relationship (Figure 6). Unweighted networks are easy to construct, but may conceal useful information.

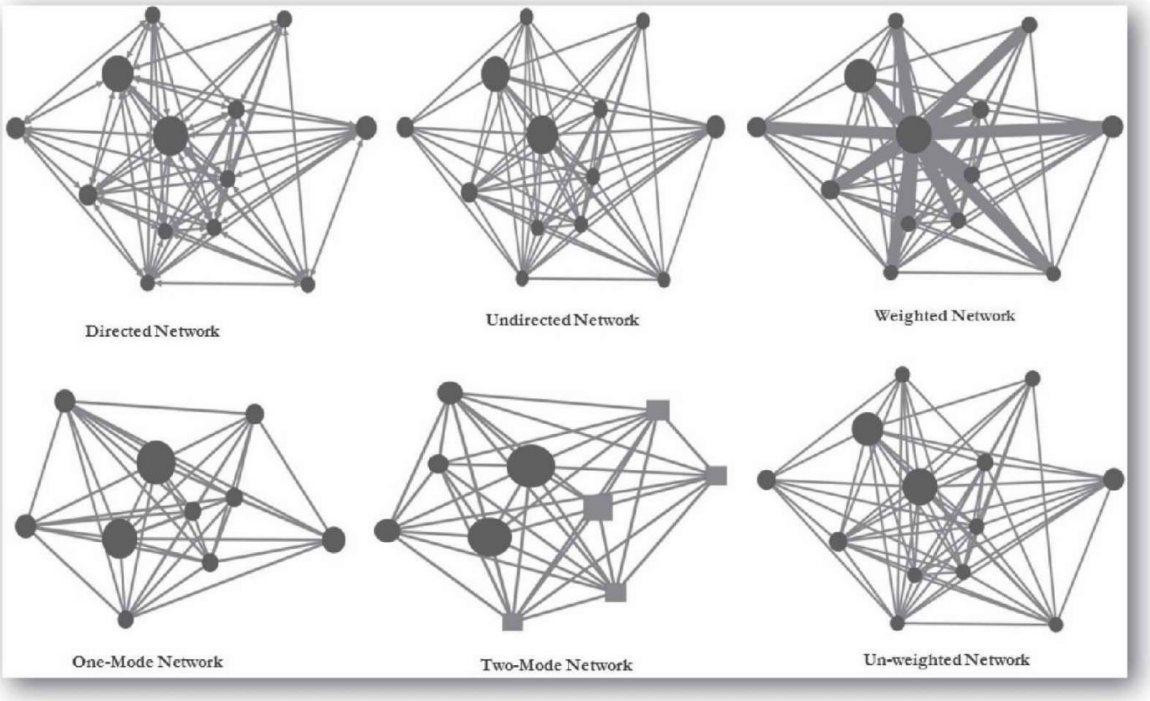

Figure 6. Types of social media networks

Keep in mind that the above-classified types are not mutually exclusive and can exist in a single network. For example, there may exist a directed weighted one-mode network. Or one could construct an undirected two-mode unweight network, and so forth.

COMMON NETWORK TERMINOLOGIES Now, let's look at some common network terminologies or properties. Network properties can be divided into two categories 1) node-level properties, and 2) network-level properties.

NODE-LEVEL PROPERTIES Node-level properties focus on one node and its position in the network. Some important node properties include degree centrality, betweenness centrality, eigenvector centrality, and structural holes.

Degree Centrality Degree centrality of a node in a network measures the number of links a node has to other nodes (Hanneman and Riddle 2005). In a Facebook network, for example, this will measure the number friendship ties a user has. In a Twitter network, it will equate to the number of followers a user has. In a directed network, degree can be either in-degree or out-degree. In-degree is the number of incoming links a node in a network receives. For example, in a Twitter network, in-degree represents the number of followers a person has. Out-degree represents that number of out links a node sends. In a Twitter network, for example, the number of people a person follows represents out-degree of a person (node). In certain networks, such as a Twitter network, in-degree (number of followers a person has) is a more important measure of a node's influence than out-degree (number of people a person follows).

Betweenness Centrality Betweenness centrality is related to the centrality (or position) of a node in a network. The nodes with high betweenness centrality have the ability to control or facilitate collaboration or flow of information due to their central position in the network (Liu, Bollen et al. 2005). In a Facebook friendship network, for example, the users who occupy the central position are better positioned to control the flow of social media content.

Eigenvector Centrality Eigenvector centrality measures the importance of a node based on its connections with other important nodes in a network. It can provide an understanding of a node's networking ability relative to that of others (Marsden 2008).

Structural Holes The idea of structural holes was first put forward by Burt (Burt 1992) who suggest that in a network exists when a certain node has an advantage or disadvantage of its location in a network (Hanneman and Riddle 2005). A node that is connected to users who are themselves not directly connected has the opportunity to mediate between them and profit from this mediation (Nooy, Mrvar et al. 2005). In a social media network, some users, because of their network position, may have an advantage or disadvantage in terms of opportunities to form and propagate information. New ideas and information mostly come from structure holes (or week ties) that exist in a network. A user with more week ties can receive novel ideas and information from remote network clusters.

NETWORK-LEVEL PROPERTIES Network properties provide insight into the overall structure and health of a network. Important network-level properties include clustering coefficient, density, diameter, average degree, and components.

Clustering Coefficient The clustering coefficient of a network is the degree to which nodes in a network tend to cluster or group together.

Density The density of a network deals with a number of links in a network. Density can be calculated as the number of links present in a network divided by the number of all possible links between pairs of nodes in a network (for an undirected network, the number of all possible links can be calculated as $n (n – 1)/2$; where n is the number of nodes in a network). A fully connected network, in which each node is connected to every other node, will have a density of 1.

Components Components of a network are the isolated sub-networks that connect within, but are disconnected between, sub-networks (Hanneman and Riddle 2005). In a connected component, all nodes are connected and reachable, but there is no path between a node in the component and any node not in the component (Wasserman and Faust 1994). The main or largest component of a network is the component with the largest number of nodes.

Diameter The diameter of a network is the largest of all the calculated shortest path between any pair of nodes in a network (Wasserman and Faust 1994), and it can provide an idea of how long it would take for some information/ideas/message to pass through the network.

Average Degree The average degree centrality measures the average number of links among nodes in a network.

NETWORK ANALYTICS TOOLS

NodeXL: NodeXL (an add-in for Microsoft Excel) is the free tool for social network analysis and visualization. It can help you construct and analyze Facebook networks (based on colikes and cocomments), Twitter networks (followers, followings, and tweets), and YouTube networks (user network and comments), among others.

UCINET: UCINET is a social network analysis software application for windows operating system. It also includes Netdraw tool for network visualization. It can be downloaded and used for free for 90 days: https://sites.google.com/site/ucinetsoftware/home.

Pajek: Pajek is a software application for analyzing and visualizing large networks (http://mrvar.fdv.uni-lj.si/pajek/). Pajek runs on Microsoft Windows operating systems and is free for noncommercial use.

Netminer: Netminer (http://www.netminer.com/) is also a software application for large social network analysis and visualization. The application can used be for free for 28 days.

Flocker: Flocker (http://flocker.outliers.es/) is a Twitter real-time retweets and mentions networks analytics tool.

Reach: Reach is an online platform to map hashtag networks and identify the most influential accounts in the Twitter conversation: http://www.reachsocial.com/.

Mentionmapp: This online tool is used to investigate Twitter mentions networks (http://mentionmapp.com/).

Subject: Social Media Analytics

Module 3: Social Media Text, Action & Hyperlink Analytics

SOCIAL MEDIA TEXT ANALYTICS

Text is one of the fundamental elements of the social media platforms. Textual elements of social media include comments, tweets, blog posts, product reviews, and status updates. Social media text analytics, also known as text mining, is a technique to extract, analyze, and interpret hidden business insights from textual elements of social media content. Organizations use text analysis techniques to extract hidden valuable meaning, patterns, and structures from the user-generated social media text for business intelligence purposes. Text analytics, for example, is useful in gaining a quick and accurate understanding of the emotion and sentiment expressed over social media channels (e.g., tweets or Facebook comments) related to a brand or a new product launch. The case study included in the chapter demonstrates this point and shows how Flyertalk.com successfully mined the textual feedback that their current and potential customers provide in their website. The volume and speed at which the comments over social media are generated does not allow for manual reading, and calls for advanced text analysis techniques. Text analytics has evolved into a well-established field with roots in variety of domains, including data mining, machine learning, natural language processing, knowledge management, and information retrieval. Studies have suggested that approximately 80 percent of data in an organization is textual in nature; in this book, however, we only focus on social media text analytics.

TYPES OF SOCIAL MEDIA TEXT Based on its nature, social media text can be broadly classified into two categories: 1) dynamic text and 2) static text.

1.DYNAMIC TEXT Dynamic text is a real-time social media user-generated text or statement to expresses an opinion about content or information posted over social media. Dynamic text is mostly posed by social media users in response to social, political, economic, personal, cultural, or business issues to express their views and feelings related to it. Dynamic text is usually smaller in length (e.g., a couple of sentences), diverse in nature, and is updated or deleted more frequently. Examples of dynamic social media text include tweets, Facebook comments, and product reviews. Below, we briefly explain the two most common dynamic social media texts: tweets and comments.

Tweet

A tweet is a one hundred forty-character massage posted by a Twitter user. A tweet may include text, images, video, or links to other websites. A tweet may also include a hashtag (#). Hashtags are used to mark keywords or topics in a tweet, and are organically created by Twitter users as a method to categorize messages. A keyword marked by a hashtag can easily appear in Twitter search, and popular hashtags are often trending topics over Twitter. Tweets accumulate over time, carry a time stamp and user information, and mostly appear in descending order; that is, the most recent first. Tweet data provides a valuable source for mining value business insights,

including exploring trending topics, measuring brand sentiment, and gathering feedback on new products and services.

Comments

Social media comments are written (usually short) statements that express opinions about content or information posted over social media. While most comments are text only, it can also include images, video, or links to other websites. The ability to post comments and participate in social media discourse is the underlying characteristic that distinguishes social media from traditional media (e.g., TV and print). Like tweets, social media comments are also a great source for mining valuable business insights from social media. Almost all social media platforms provide commenting features. Comments accumulate over time, carry a time stamp and user information, and mostly appear in descending order; that is, the most recent first.

Discussion

Discussion takes the form of textual or written conversation or debate about a certain topic, product, or service. Mostly, discussions among social media users happen through Internet forums. Internet discussion forums are treelike in structure; that is, a forum can contain a number of sub-forums focused on specific topics or threads. In these forums users can post questions and reply to questions posted by other users. Discussions accumulate over time, carry a time stamp and user information, and mostly appear in descending order; that is, the most recent first. Vault Network is an example of an Internet forum that focuses on online games.

Conversation

Social media textual conversation (also known as chatting) is an instant exchange of short written messages between two more people. Chatting usually takes a casual form and are carried out through dedicated messaging services/tools. A variety of messaging tools have been developed for textual conversation, including desktop-based (e.g., Skype); web-based (e.g., Google Hangouts and Facebook chat) and mobile-based (e.g., Viber). Note that these services are not only limited to textual conversation, but also support video and voice conversation. And now that all media are converging, most of the messaging services also come in desktop, mobile, and web forms. For example, Skype has both desktop and Smartphone versions. An important point to note here is that most of the social media textual conversation is private in nature and may not be subject to mining.

Reviews

Reviews are critical evaluations of a product or service performed by customers and experts. They can take both longer and shorter forms. Reviews by customers are mostly shorter when compared to formal reviews by experts. Reviews can include textual elements and ratings. ProductReview.com.au, for example, is a site devoted to product/service reviews and ratings

submitted by customers. Product reviews can serve as an excellent source for mining customers' opinions and feelings about a product or service.

2.STATIC TEXT Static social media text is usually large in length (e.g., several paragraphs) and is generated, updated, or deleted less frequently. Examples of static text include wiki content, a blog page, Word documents, corporate reports, electronic mail (e-mail), and news transcripts. At the highest level of abstraction, the purpose of static social media text is to inform, educate, and elaborate.

PURPOSE OF TEXT ANALYTICS

Both dynamic and static text are subject to analytics. The following are some of the objectives of social media text analytics for business intelligence purposes (Figure 4).

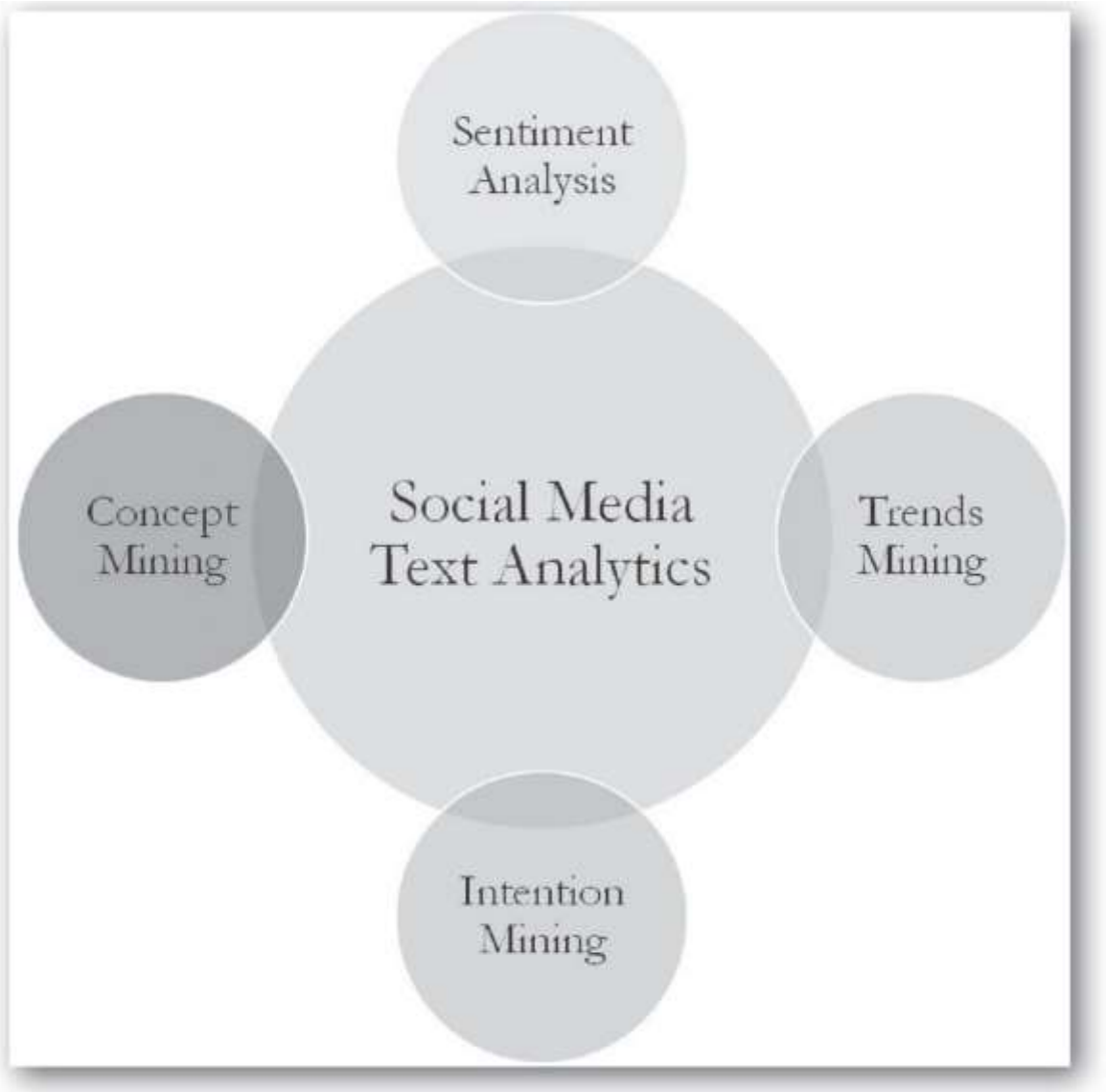

Figure 4. Purpose of social media text analytics

SENTIMENT ANALYSIS

Sentiment analysis analyzes and categorizes social media text as being positive, negative, or neutral. Social media sentiment analysis mostly focuses on dynamic text. The primary purpose of sentiment analysis is to determine how your customers feel about a particular product, service, or issue. For example, as a product manager, you might be interested to know how your customers on Twitter feel about a product/service that was recently launched. Analyzing your tweets or Facebook comments may provide an answer to your question. Using sentiment analysis, you may be able to extract the wordings of the comments and determine if they are positive, negative, or neutral. At the end of the chapter, several analytical tools are listed for semantic analysis. A later section in the chapter provides a step-by-step guide on analyzing social media text using Semantria.

Semantria is an example of a text sentiment analysis tool. It will go through the following steps to extract sentiments from a document:

Step 1: It breaks the document into its basic parts of speech, called POS tags, which identify the structural elements of a sentence (e.g. nouns, adjectives, verbs, and adverbs).

Step 2: Algorithms identify sentiment-bearing phrases like "terrible service" or "cool atmosphere."

Step 3: Each sentiment-bearing phrase earns a score based on a logarithmic scale ranging from negative ten to positive ten.

Step 4: Next, the scores are combined to determine the overall sentiment of the document or sentence. Document scores range between negative two and positive two. For example, to calculate the sentiment of a phrase such as "terrible service," Semantria uses search engine queries similar to the following:

"(Terrible service) near (good, wonderful, spectacular)"
"(Terrible service) near (bad, horrible, awful)"

Each result is added to a hit count; these are then combined using a mathematical operation called "log odds ratio" to determine the final score of a given phrase.

INTENTION MINING

Intention or intent mining (Chen, Lin et al. 2002) aims to discover users' intention (such as buy, sell, recommend, quit, desire, or wish) from natural language social media text such as user comments, product reviews, tweets, and blog posts. Social media as the integral part of our contemporary lives and is widely used by millions of customers to express desires, needs, and intention (Niven 2013). Companies may use intent mining to find new potential customers who intend to buy a product (or services) and service existing customers who have trouble with a product. For example, an analysis of company-related tweets may detect purchase intention based on the presence of the word "buy" or "purchase." Similarly, detecting the "quit" intention

may identify and service the customers at risk of leaving the company. The Semantria analytical tool discussed later in this chapter, for example, can be used to mine intentions.

TRENDS MINING

Trends mining, also known as predictive analytics, uses huge amounts of historical and real-time social media data to predict future events. For example, a vast amount of social media data (e.g., comments and tweets) can be mined to identify patterns and trends for new product or service development or to improve customer satisfaction by anticipating their needs. Trend mining exploits patterns in large amounts of data by using sophisticated statistical techniques, including machine learning, data mining, and social network analysis. Predictive analysis using conventional business data has been used in a variety of domains, including marketing, banking, telecommunication, and healthcare. However, social media predictive analytics is still an emerging practice and may take some time for sophisticated tools and techniques to emerge.

CONCEPT MINING

Concept mining aims to extract ideas and concepts from documents. Unlike text mining, which is focused on extracting information, concept mining extracts ideas from large document sets. Thus, concept mining is useful in extracting ideas from large amounts of static social media text, such as wiki content, a web page, Word documents, and news transcripts. Concept mining can be employed to classify, cluster, and rank ideas.

STEPS IN TEXT ANALYTICS

Text analytics, like any other form of social media analytics, is the art and science of getting the desired business intelligence from the text posted over social media (Figure 5). While the steps required for text analytics are largely dependent on the type of approach and tool employed, a typical social media text analysis includes the following cyclical steps.

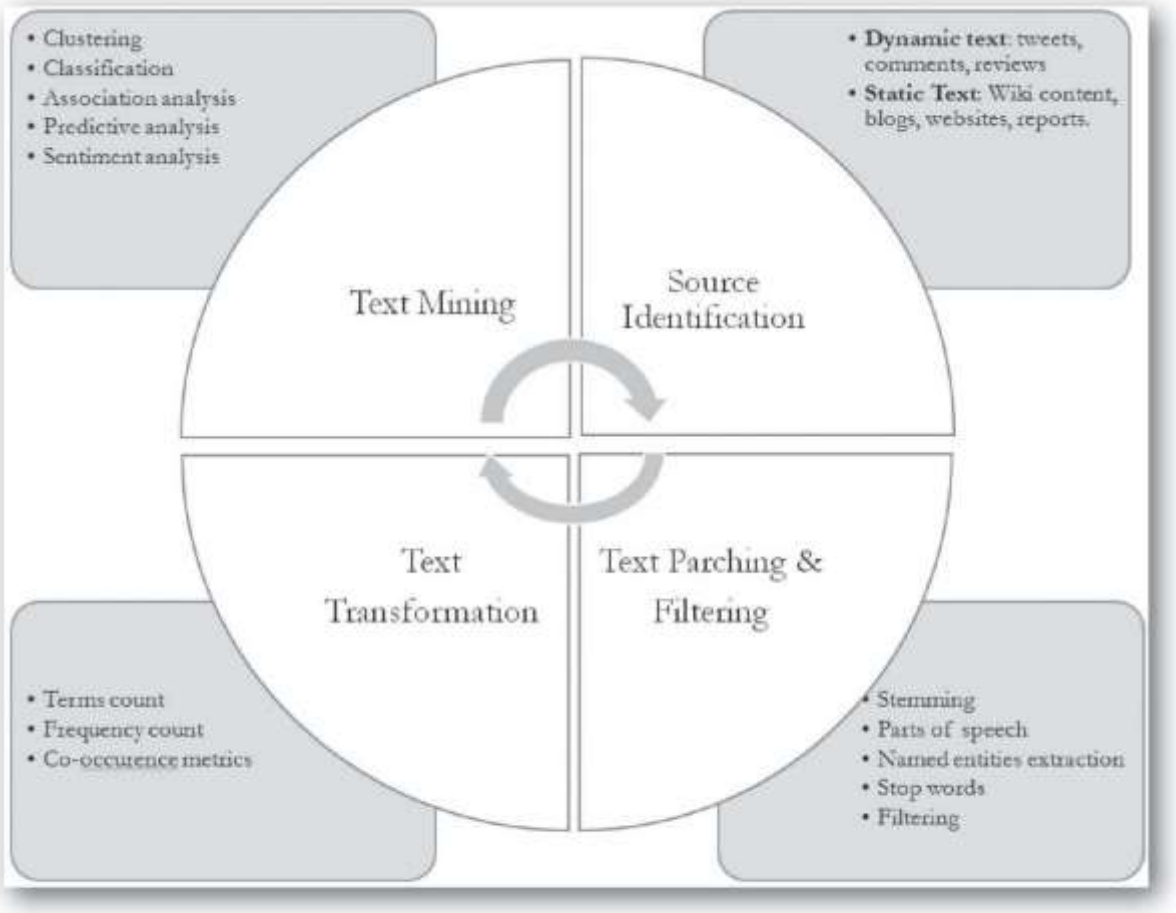

Figure 5. Steps in text analytics

1.IDENTIFICATION AND SEARCHING

The text analytics process starts with identifying the source of the text that will be analyzed. Text posted on social media is dynamic, huge, diverse, multilingual, and noisy. Thus, finding the right source for the purpose of text analytics is very crucial for gaining useful business insights. The genre of the source text also will determine the type of tool used to extract and analyze it. For example, extracting tweets requires different tools and approaches than analyzing a document or website text. Analyzing tweets, for example, requires API-based searching and extraction of data from the Twitter timeline based on criteria that you specify. You can choose to extract tweets that include specific keywords, such as your company name. The desired business question that needs to be answered with text analytics will serve as a good starting point.

2.TEXT PARSING AND FILTERING

The next step is to parse, clean, and filter the text, and create a dictionary of words using NPL, which is mostly based on machine learning techniques. In order for computer and algorithms to extract meanings from the text, the sentence structures and parts of speech are determined, named entities extracted (people, organizations, product/service names, etc.), stop words

removed, and spellings are checked. Most of these steps are automatic; however, in certain stages, human intervention is required. For example, in the filtering stage, manually cleaning (by humans with domain expertise) may be required to remove unwanted or irrelevant terms.

3.TEXT TRANSFORMATION

For analytical algorithms to be applied to the text, it should be transformed into a computer-readable format (e.g., 0s and 1s) for analysis. The cleaned text is thus transformed into numerical representations using linear algebra-based techniques, such as latent semantic analysis and vector space models.

4. TEXT MINING

At this step, the text is actually mined to extract the needed business insights. Varieties of text mining algorithms are applied to the text, such as clustering, association, classification, and predictive analysis, and sentiment analysis. Text analysis employs these sophisticated algorithms to extract sentiment and meanings from the text in a similar manner to the way human do; however, the process is thousands of times faster.

Association—Association or association mining is a data-mining technique used to determine the probability of the cooccurrence of items in a collection of documents. The relationships between cooccurring items are expressed as association rules. In text analytics, for example, social media text can be clustered together based on cooccurrence frequency. Or it can be used, for example, to find that a user who liked a social media content A and B is 90 percent likely to also like content C.

Clustering—Clustering or cluster analysis groups objects based on similarity in non-overlapping groups. Clustering is an important part of data mining and text analytics. Social media text (such as tweets or comments), for example, can be clustered into positive, negative, and natural categories. Or nodes in a social media network can be clustered based on importance.

Classification—From the text analytics perspective, classification or categorization is used to find similarities in the document and groups them with predefined labels based on the themes contained in the document (Chakraborty, Pagolu et al. 2013). For example, an e-mail can be classified as spam based on its contents.

SOCIAL MEDIA TEXT ANALYSIS TOOLS

A variety of social media text analysis tools are available on the market. Some are free and others are paid. Below we list some popular text analysis tools.

Discovertext: Discovertext (http://discovertext.com/) is a powerful platform for collecting, cleaning, and analyzing text and social media data streams.

Lexalytics: Lexalytics (http://www.lexalytics.com/) is a social media text and semantic analysis tool for social media platforms, including Twitter, Facebook, blogs, etc.

Tweet Archivist: Tweet Archivist (https://www.tweetarchivist.com/) is focused on searching, archiving, analyzing, and visualizing tweets based on a search term or hashtag (#).

Twitonomy: Twitonomy (https://www.twitonomy.com/) is a Twitter analytics tool for getting detailed and visual analytics on tweets, retweets, replies, mentions, hashtags, followers, etc.

Netlytic: Netlytic (https://netlytic.org) is a cloud-based text and social network analytics platform for social media text that discovers social networks from online conversations on social media sites.

LIWC: Linguistic Inquiry and Word Count (LIWC) is a text analysis tool for analyzing emotional, cognitive, structural, and process components present in individuals' verbal and written speech samples: http://www.liwc.net/

Voyant: Voyant (http://voyant-tools.org/) is a web-based text reading and analysis. With Voyant, a body of text can be read from a file or directly exported from a website.

SOCIAL MEDIA ACTIONS ANALYTICS

Actions are the cash cow of social media. It is what the users do on social media that matters most to social media marketers. Typical actions performed by social media users include likes, dislikes, shares, views, clicks, tags, mentions, recommendations, and endorsements. Actions are way to express symbolic reactions to social media content. Symbolic actions are an easy and fast way to express feelings, unlike written reaction in the form of textual comments. Actions are not just symbolic reactions; they carry emotions and behaviors that can be harnessed. More importantly, social media actions are social expressions; that is, a user who performs an action (e.g., liking certain content) is visible to (or shared with) other social media users, in particular with their friends. This shareable nature of social media actions makes it very attractive to social media marketers and businesses. Take as an example Moviefone (an American-based movie listing and information service company), which enabled logins with Facebook and Twitter credentials. Enabling such login services not only allow users to use the Moviefone service conveniently, but also let them connect with their social media friends and share content over the Moviefone site. Enabling social logins led to a 300 percent increase in site traffic, a 40,000 to 250,000 increase in referrals per month, and a 40 percent increase in click-through rate (Petersen 2012).

WHAT IS ACTIONS ANALYTICS?

Social media actions analytics deals with extraction, analysis, and interpretation of the insights contained in the actions performed by social media users. Social media actions are of great value to social media marketers because of their role in increasing revenue, brand value, and loyalty. Organizations can employ actions analytics to measure popularity and influence of a product, service, or idea over social media. For example, a brand marketer might be interested to know how popular their new product is among social media users. Analyzing your Facebook likes and Twitter mentions, for example, may provide an answer to your questions.

COMMON SOCIAL MEDIA ACTIONS

Below, we briefly discuss some of the most prevalent social media actions. All these actions are performed by social media users and can become your social media metrics. Metrics, in simple words, are anything you want to measure. Social media users can come in many forms, including followers, fans, and subscribers.

LIKE Like or "Like" buttons or like options are a feature of social media sites (e.g., social networks, blogs, and websites) that allow users to express their feelings of liking certain products, services, people, ideas, information, places, or content. They are actions performed by social media users to express symbolic positive reaction to social media content. Facebook's "Like" button enables users to easily express their feelings and give your product or service a virtual thumbs up. Incorporating a "Like" button in social media platforms and websites is becoming a norm. Social media platforms display accumulated likes received by content over time. Facebook's "Like" button is the most famous one. Google+ social networking platform uses a "+1" symbol to express liking. Companies use Google+ and Facebook fan pages to receive likes from customers, but the "Like" button can also be incorporated into a company website or blog. The "Like" button can be easily incorporated into a website as follows:

1. Visit Facebook's developers' page: https://developers.facebook.com/docs/plugins/like-button
2. Customize and generate a code.
3. Paste the code into your website after the tag.

DISLIKE "Dislike" buttons are included in some social media platforms (e.g., YouTube) and allow users to express their negative feelings of disliking certain content (e.g., products, services, people, ideas, information, or places) posted over social media. Similar to the "Like" feature, it is visible to others and accumulated over time. The "Dislike" button is not as prevalent as the "Like" button.

SHARE Share or "Share" button or sharing is a feature that allows social media users to distribute the content posted over social media to other users. For example, the Facebook "Share" button lets users add a personal message and customize who they share the content with. The WordPress (a blogging platform) "Share" button, for example, allows users to share their blog content across a range of social media platforms. Companies incorporate share buttons into website to boost their website traffic by channeling visitors from social media sites.

VISITORS, VISITS, REVISITS A visitor is a person who visits your website or blog. A single visitor may visit a page or content one or more times (revisits). Visits are also known as sessions. Other related concepts are: Unique visitor—A person who arrives at your page first time. Average bounce rate—the percentage of visitors who visit a website and leave the site quickly without viewing other pages. Session duration—The average duration of a visit or session.

VIEW Views are the number of times social media content (a post, video, graphic, etc.) is viewed by users. A slightly different but related concept is page views, which is each time a visitor views a page on your company website or blog.

CLICKS Clicks are the actions performed by users by pressing or clicking on the hyperlinked content of your website or blog. Through clicks, users navigate the web. Click data can be

harvested for business intelligence purposes, such as, to reduce bounce rate and improve website traffic. A technique called clickstream analysis is used by business managers for a variety of business intelligence purposes, including website activity, website design analysis, path optimization, market research, and finding ways to improve visitor experience on the website. The clickstream is the semistructured data trail/log (such as date and time stamp, IP address, and the URLs of the pages visited) a user leaves while visiting a website.

TAGGING Tagging is the act of assigning or linking extra pieces of information to social media content (such as photographs and bookmarks) for identification, classification, and search purposes. Tagging lets user classify social media content the way they see it. Tagging may take a variety of forms. For example, bloggers can attach descriptive keywords (tags) to their posts to facilitate classification and searching of content, and Facebook users can add tags to anything they post on their status, including photos and comments. Social bookmarking services (such as del.icio.us) let users organize their bookmarks flexibly by adding descriptive tags. This practice of collaborative tagging is commonly known as folksonomy—a term coined by Thomas Vander Wal (Wal 2005). These days, almost all prominent companies (e.g., Facebook and Flickr) provide tagging services to their users. Because the contents are tagged with useful keywords, social tagging expedites the process of searching and finding relevant content.

MENTIONS Mentions or social mentions are the occurrence of a person, place, or thing over social media by name. For example, a brand name maybe mentioned in a Facebook comment, blog post, YouTube video, or tweet. Mentions are important and can indicate popularity of person, place, or thing. For example, a social marketer may be able to gauge the popularity of a product/service/campaign by mining Twitter mentions data. A Twitter mention is the inclusion of a "@username" in a tweet.

HOVERING Hovering is the act of moving a cursor over social media content. Capturing users' cursor movement data can help you understand user behavior on a social media site. Cursor movement/hovering over an ad, for example, can be considered as a proxy for attention. Most people who view an ad do not necessary click on it, thus if we are relying on clicks analytics only, we may lose a vital piece of information (i.e., attention). Studies have even suggested a strong correlation between hover time and purchases. Traditionally, hovering data has been used in website design and for improvement of user experience.

CHECK-IN Check-in is a social media feature that allows users to announce and share their arrival at a location, such as a hotel, airport, city, or store. Many social media services, including Facebook and Google+, provide check-in features. The location of the user is determined using GPS (global positioning system) technology. Check-in data can, for example, be mined to offer location-based services/products.

PINNING Pinning is an action performed by social media users to pin and share interesting content (such as ideas, products, services, and information) using a virtual pinboard platform. Some famous pinning platforms include Pinterest, Tumblr, StumbleUpon, or Digg. Business can use these virtual pin boards to share information and connect with and inspire their customers. Four Seasons Hotels and Resorts, for example, use Pinterest to curate travel, food, and luxury lifestyle content to inspire customers.

EMBEDS Embedding is the act of incorporating social media content (e.g., a link, video, or presentation) into a website or blog. An embed feature lets users embed interesting content into their personal social media outlets.

ENDORSEMENT Endorsement is a features of social media that lets people endorse and approve other people, products, and services. For example, LinkedIn lets user endorse the skills and qualifications of other people in their network.

UPLOADING AND DOWNLOADING In simple words, uploading is the act of adding new content (e.g., texts, photos, and videos) to a social media platform. The opposite of uploading is downloading; that is, the act of receiving data from a social media platform. All most all social media content is created and uploaded by users, which is better known as user-generated content. For some companies, uploading and downloading is the single most important action to measure. For Instagram and Flickr, which are both photo-sharing platforms, the number of photos uploaded daily matters more than anything else.

ACTIONS ANALYTICS TOOLS

Currently, there is no single platform that can capture all the actions discussed in this chapter. Certain platforms can be employed to measure social media actions across platforms. Below we list some popular actions analytics tools.

Hootsuite: Hootsuite is an easy-to-use online platform that enables you to manage your social media presence across the most popular social networks. Hootsuite offers different plans depending on your business needs and budget: free, pro, or enterprise. In this tutorial, we will employ the free version, which supports up to five social media profiles and has limited analytics information. **SocialMediaMineR:** SocialMediaMineR is a social media analytics tool that takes one or multiple URLs and returns the information about the popularity and reach of the URL(s) on social media, including the number of shares, likes, tweets, pins, and hits on Facebook, Twitter, Pinterest, StumbleUpon, LinkedIn, and Reddit. The tool can accessed from here: http://cran.rproject.org/web/packages/SocialMediaMineR/index.html

Lithium: Lithium (http://www.lithium.com/) is social media management tool that provides a variety of products and services, including social media analytics, marking, crowd-sourcing, and social media marketing.

Google Analytics: Google Analytics (http://www.google.com/analytics/) is an analytical tool offered by Google to track and analyze website traffic. It can also be used to for blogs and wiki analytics. **Facebook Insights:** Facebook Insights (https://www.facebook.com/insights/) helps Facebook page owners understand and analyze trends within user growth and demographics.

Klout: Klout (https://klout.com/) measures your influence across a range of social media channels based on how many people interact with your posts. Your Klout score measures your influence on a scale from one to one hundred.

Topsy: Topsy (http://topsy.com/) is similar to Icerocket and Social Mention, with its main focus around social media, especially multimedia sites and blogs.

Tweetreach: This tool helps you measure the number of impressions and reach of hashtags. The tool can be accessed here: https://tweetreach.com

Kred: Kred helps you measure the influence of a Twitter account: www.kred.com

Hashtagify: This tool measures the influence of hashtags: http://hashtagify.me

Twtrland: Twtrland is a social intelligence research tool (http://twtrland.com/) for analyzing and visualizes your social footprints.

Tweetstats: using your Twitter user name, Tweetstats graphs Twitter stats including tweets per hour, tweets per month, tweet timelines, and reply statistics (http://www.tweetstats.com).

SOCIAL MEDIA HYPERLINK ANALYTICS

Hyperlinks are the pathways of social media traffic. Hyperlinks are references to web resources (such as a website, document, and files) that users can access by clicking on it. Hyperlinks can link resources within a document (inter-linking) and among documents (intralinking). For example, clicking on a hyperlink in a tweet can link you to other resources (e.g., websites) available over the Internet. Hyperlinks are not merely technical links between two websites, but serve a more symbolic means (Park 2003; Kim and Nam 2012). As a website is an official and unique entity representing an organization itself (Garrido 2003); therefore, embedding hyperlinks in an organization's website can be considered an official act of communication between two organizations. Hyperlinks among websites represent not only a reasonable approximation of a social relationship (Jackson 1997), but also serve as a symbolic meaning of validating or endorsing the linked organization (Vreelnad 2000). In conjunction with this, these hyperlinks that exist between two organizational websites reflect a sense a sense of validation, trust, bonding, authority, and legitimacy (Vreelnad 2000; Park 2003; Nam, Barnett et al. 2014). Websites mostly connect or link to other websites of similar nature, so hyperlinks can also serves as indicators of content similarity (Chakrabarti, Joshi et al. 2002).

TYPES OF HYPERLINKS From hyperlink analytics point of view, mainly there are three types of hyperlinks, 1) in-links, 2) out-links, and 3) co-links.

IN-LINKS In-links are the incoming hyperlinks or links directed toward a website or originated in other websites (Björneborn and Ingwersen 2004). For example, consider the top left image in the Figure 9, page A is receiving two in-links coming from pages B and C.
In-links are of great interest to social markers, because they bring traffic to a particular website. Thus, harvesting them can help us understand where the traffic to a corporate website is coming from. In-links also play an important role in website analytics, as both the quality and number of in-links can impact the search engine ranking of the website (more details on this are provided in the search analytics chapter). (Thelwall 2001) In-links can also impact the popularity of social media contents. A study on YouTube viral videos, for instance, found that among other things, in-links play crucial roles in the viral phenomenon, particularly in increasing views of videos posted on YouTube (Khan and Vong 2014). Studies have also shown that in-link counts strongly correlate with measures describing business performance (Vaughan 2004).

OUT-LINKS Out-links are hyperlinks generated out of a website (Bjorneborn 2001). As shown in the top-right image in the Figure 9, page A is sending two out-links: one to page B and one to page C. **CO-LINKS** Co-links have two dimensions. First, if two websites receive a link from a third website, they are considered to be connected indirectly. For example, page A links to both pages B and C, therefore B and C are considered to be co-linking, or connected indirectly (bottom-left image in the Figure 9). Second, if two pages link to a third page, they are also considered to be colinking. As shown in the bottom-right corner of the Figure 9, Pages B and C are linking to page A; therefore, B and C are connected indirectly. Co-links have been used to compare and map competitive similarity among companies (Vaughan and You 2006).

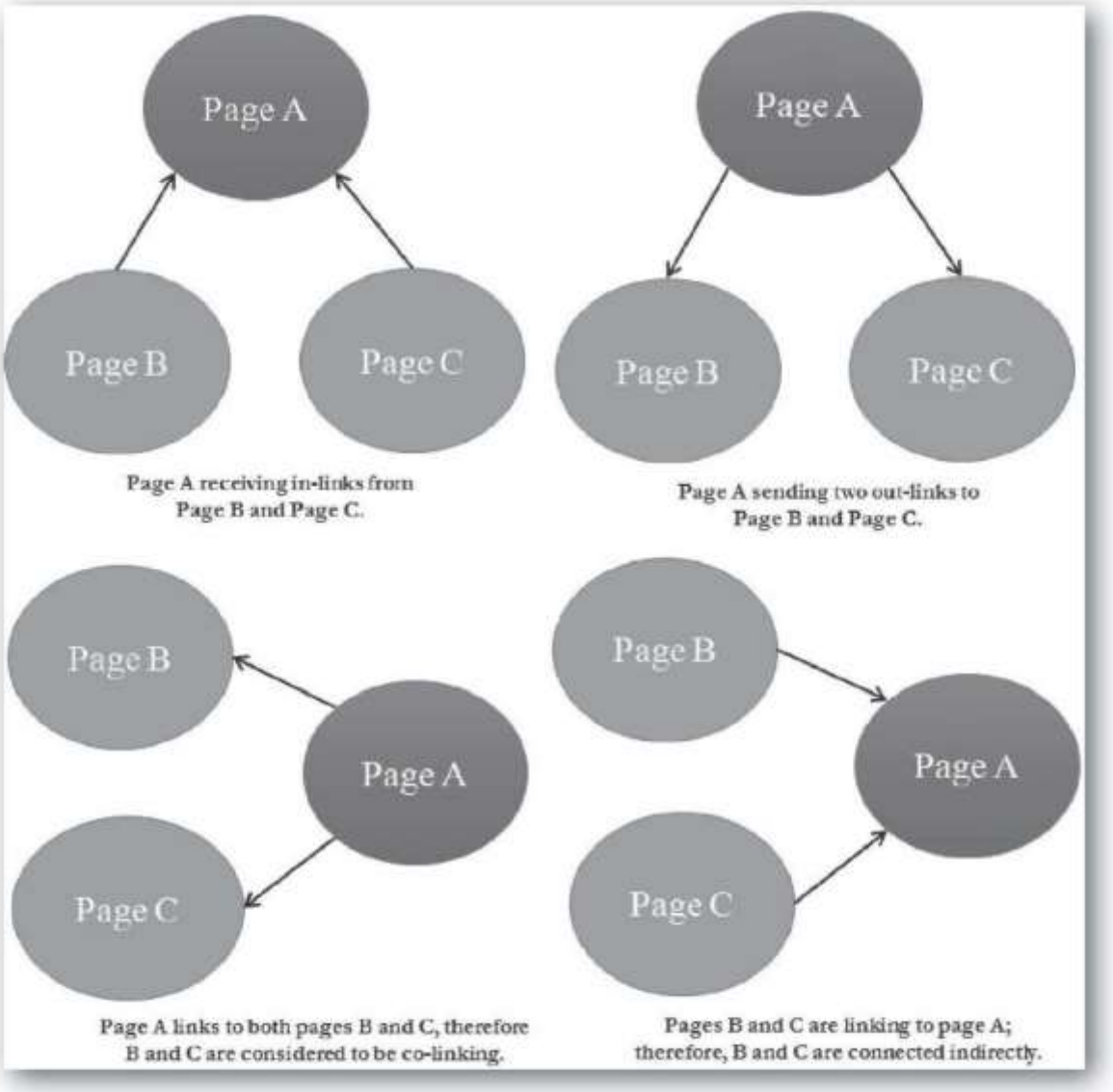

Figure 9. Different types of hyperlinks

HYPERLINK ANALYTICS

Hyperlink analytics deals with extracting, analyzing, and interpreting hyperlinks (e.g., in-links, out-links, and co-links). The basic assumption pertaining to hyperlink analytics is that the number and quality of hyperlinks to a website equates to its importance or value (Thelwall 2014). Hyperlink analytics can also reveal, for example, Internet traffic patterns and sources of the incoming or outgoing traffic to and from a website. Hyperlink analysis has been used to study a variety of topics including ranking of universities, understanding the blogosphere, scholarly websites (Vaughan and Thelwall 2003), and political networks (Park and Thelwall 2008), and to measure business competitiveness (Vaughan and You 2006). The case study included in this chapter demonstrates the importance of hyperlinks in viral phenomena and shows the valuable insights they carry for viral marketers in formulating viral marketing strategies. We must admit that, regardless of its importance, hyperlink analysis also has some limitations. Hyperlink networks, for example, does not provide any insight into the type or amount of traffic flowing among websites(Ackland 2010).

When we talk about hyperlinks analytics, it mostly implies in-links, outlinks, and co-links analysis and does not include hyperlinks within a website between pages. Hyperlinks between pages within a website are created mostly for navigational purposes. Also, search engine ranking algorithms either ignore or give low importance to hyperlinks within a website (Thelwall 2014).

TYPES OF HYPERLINK ANALYTICS Hyperlink analytics can take several forms, including: 1) hyperlink environment analysis, 2) link impact analysis, and 3) social media hyperlink analysis.

1.Hyperlink Environment Analysis
Hyperlink environment analyses deal with a particular website or set of websites. Hyperlinks (i.e., out-links, in-links, and co-links) of a website are extracted and analyzed to identify the sources of Internet traffic. Hyperlinks environment networks can take two forms: 1) co-links networks or 2) in-links and out-links networks.

Co-Link Networks In co-links environment networks, nodes are websites and links that represent similarity between websites, as measured by co-link counts. With the Webometric Analyst tool, one can construct a co-link network diagram among a set of websites (Thelwall 2005; Thelwall 2014).

In-Links and Out-Links Networks In-links and out-links hyperlink environment networks are constructed based on in-links and out-links from a website or set of websites. In such a network, nodes will be websites and links will present in-links and out-links. The tutorial provided in this chapter demonstrates constructs such as network using the VOSON hyperlink analysis tool.

2.Link Impact Analysis
Link impact analysis investigates the web impact of a website address (or URL) in terms of citations or mentions it receives over the web. In a link impact analysis, statistics about web pages that mention the URL of a given website are collected and analyzed (Thelwall 2005; Thelwall 2014). The assumption is that a URL (or website address) cited frequently over the web is more important. Thus, measuring the web impact of URLs may provide an idea about the importance of a website.

3.Social Media Hyperlink Analysis
Social media hyperlink analysis deals with extraction and analysis of hyperlinks embedded within social media texts (.e.g., tweets and comments). These hyperlinks can be extracted and studied to identify the sources and destination of social media traffic. A good example of the usefulness of the hyperlink embedded in the social media text is the study by Khan et al. (2014), in which they extracted out-links from Korean and US government agencies' tweets. By extracting out-links and tracing them back to their sender, the authors were able to constructed a map of the out-link structure (Figure 10). According to a comparison of out-links between tweets of the Korean and US governments, there were some differences in citation (i.e., out-link) patterns. The Korean government tended to cite domestic portals' news services and their own blogs (i.e., self-citation). Although there were SNSs and newspaper sites, most of the related out-links were for portals. On the other hand, the US government showed a more diverse pattern in terms of out-link destinations. US out-links were not concentrated in specific sites

and tended to go directly to news agencies, not to secondary sources such as portals. These comparisons between the US and Korean governments suggest that social media out-links can carry valuable information and can help explain real-world phenomena and shed light on the disparities in social media use among different cultures (Khan, Yoon et al. 2014).

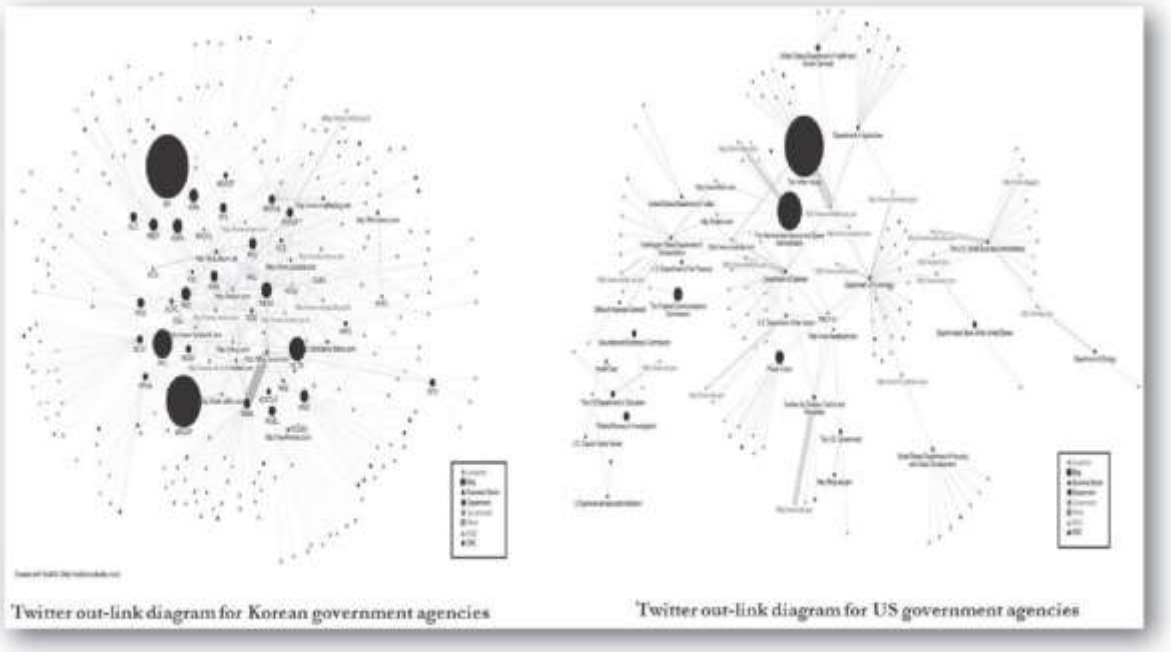

Figure 10. Twitter networks for Korea (left) and the US Governments (right)

HYPERLINK ANALYTICS TOOLS

The following are some popular hyperlink analytics tools.

Webometric Analyst: Webometric Analyst is a web impact analysis tool and can conduct variety of analysis on social media platforms including hyperlink network analysis and web mentions: http://lexiurl.wlv.ac.uk/

VOSON: VOSON (http://www.uberlink.com/) is a hyperlink analytics tools for constructing and analyzing hyperlink networks. More details on VOSON are provided in the hyperlink analytics chapter. This chapter includes a detailed tutorial on using VOSON for hyperlink analysis.

Open Site Explorer: Open Site Explorer is a link analysis tool to research and compare competitor backlinks, identify top pages, view social activity data, and analyze anchor text: https://moz.com/researchtools/ose/

Link Diagnosis: Link Diagnosis (http://www.linkdiagnosis.com/) is a free online tool for analyzing and diagnosing links.

Advanced Link Manager: Advanced Link Manager provides a variety of link analysis capabilities, including the ability to track link-building progress over time, domain quality analysis, backlinks evolution, and website-crawling abilities: http://www.advancedlinkmanager.com/

Majestic: Majestic (https://majestic.com) provides a variety of link analysis tools, including link explorer, backlinks history, and link mapping tools.

Backlink Watch: Backlink Watch (http://backlinkwatch.com/) is a free tool for checking the quality and quantity of in-links pointing to a website.

Subject: Social Media Analytics

Module 4: Social Media Location & Search Engine Analytics

LOCATION ANALYTICS:

Location Analytics, also known as spatial analysis or geo-analytics is concerned with mapping, visualizing, and mining the location of people, data, and other resources. All sectors, including business, government, non-profit, and academia, can benefit from location analytics. Thanks to the GPS (global positioning systems) embedded in mobile devices, providing location-based services, products, and information is becoming a reality.

SOURCES OF LOCATION DATA:

Location information can come from a variety of sources, including the following:

POSTAL ADDRESS: Most business analytics applications rely on address information of their customers, including city names, locality names, and postal or zip codes.

LATITUDE AND LONGITUDE: In geography, latitude (shown as a horizontal line on a globe) and longitude (shown as a vertical line on a globe) are used to find exact location on Earth.

GPS-BASED GPS: is a satellite-based navigation system that can be used find exact location people and resources. Mobile analytics mostly rely on GPS-based location data. GPS-based location analytics can provide us the most accurate location for social media users.

IP-BASED: Public IP (Internet protocol) can be used to determine the location of Internet users. A public IP address is an exclusive numerical address (like a home address) assigned to a device connected to the Internet. Different regions in the world are assigned a specific block of public IP addresses; hence, it can be used to mine approximate geo-location of Internet users.

CATEGORIES OF LOCATION ANALYTICS:

Based on its scope, location analytics can be broadly classified into two categories: 1) business data-driven location analytics, 2) social media data driven location analytics

1. **BUSINESS DATA-DRIVEN LOCATION ANALYTICS:**
   Business data-driven location analytics deals with mapping, visualizing, and mining location data to reveal patterns, trends, and relationships hidden in tabular business data. Capitalizing on the data stored in a business database, location analytics, for example, can map and capture vast among of geo-specific data to provide information, products, and services based on where customers are. Using the location of customers, for instance, it is possible to recommend the nearest convenience store, coffee shop, taxi, or even probable social relations. Or it can be used for any other business decision, such as, what is the best potential new site for a business warehouse?

Applications of Business Data-Driven Location Analytics:

Business data-driven location analytics has several applications, including the following:

- **Powerful Intelligence** Simple maps have been widely used, but they are limited in providing insightful details. Using sophisticated mapping techniques, such as clustering, heat mapping, data aggregation (e.g., aggregating data to regions), and color-coded mapping, can generate powerful business intelligence.

- **Geo-Enrichment** Simple data maps can be enriched with customer data, including demographic, consumer spending, lifestyle, and locations. For example, where do my loyal customers spend most of their time?

- **Collaboration and Sharing** Maps are easy to understand and are good communication and collaboration tools. Location analytics can map business data for collaboration across organization. It can also be used for information sharing purposes with customers. At end of this chapter, a step-by-step tutorial is provided to map sample tabular business data using Google Fusion Tables. With Google Fusion Tables, you can map data and display and share the results as maps, tables, and charts.

2. SOCIAL MEDIA DATA-DRIVEN LOCATION ANALYTICS

Social media data-driven analytics relies on social media location data to mine and map location of social media users, content, and data. Social media location information comes mainly from GPS and IP.

Uses of Social Media-Based Location Analytics Social media location–based services are becoming a day-to-day reality. Organizations use location-based services for a variety of purposes, including the following:

1. **Recommendation Purposes**: Organizations can harvest location data to recommend products, services, and social events to potential customers in real time as they approach certain localities. For example, Tender recommends potential social relationships based on the location of users.
2. **Customer Segmentation**: Social media location data can be used to segment customers based on their geographic location. Tweepsmap (https://tweepsmap.com/), for example, can be used to geo-locate your Twitter followers by country, state, or city.
3. **Advertisement**: Location-based advertisement allows targeted marketing and promotion campaign mostly delivered through mobile devices to reach specific target audiences.
4. **Information Request**: Based on their current location, customers can request a product, service, or resource (e.g., the nearest coffee shop, restaurant, or parking lot).

5. **Alerts**: Location data can be used to send and receive alerts and notifications, such as sales and promotion alerts traffic congestion alerts, speed limit warnings, and storm warnings.
6. **Search and Rescue**: Location data is vital in search and rescue operations. For example, Agos, a geotagging and reporting platform that enables communities deal with climate change adaptation and disaster risk reduction.
7. **Navigation**: Mobile-and GPS-based navigation services and apps assist us in finding addresses. BE-ON-ROAD, for instance, is a free offline turn-by-turn GPS navigation app for Android devices.

LOCATION ANALYTICS AND PRIVACY CONCERNS

While location-based services bring ease, convenience, and safety to customers and value to business, they also raise serious privacy issues related to collection, retention, use, and disclosure of location information (Minch 2004). Tracking, mining, and storing location information can endanger some fundamental human rights, such as freedom of movement and freedom from being observed. Minch (2004) raised several issues arising from location-based services, including the following.

- Should users of location-enabled devices be informed when location tracking is in use?
- Should users of location-enabled devices be permitted to control the storage of location information?
- Should location information as stored be personally identifiable, or should the user have the option to preserve degrees of anonymity?
- What legal protection should a person's historical location information have against unreasonable search and seizure?
- To what extent should users of location-based services be allowed to choose their own level of identifiability/anonymity?
- What level of disclosure control should be dictated by government regulation? By the affected individual customers, users, etc.? By other parties?
- What governmental legislation and regulation is appropriate to assure citizens' rights of privacy in an era of location-aware mobile devices?

LOCATION ANALYTICS TOOLS

Google Fusion Tables: Google Fusion Tables is a web service to geo-tag, store, share, query, and visualize tabular business data overlaid on Google Maps.

Agos: Agos is a geo-tagging and reporting platform that helps communities deal with climate change adaptation and disaster risk reduction: http://agos.rappler.com/#

Tweepsmap: Tweepsmap (https://tweepsmap.com/) maps your Twitter followers by country, state, or city.

Trendsmap: Trendsmap (http://trendsmap.com/) is real-time tool that maps the latest trends from Twitter, anywhere in the world.

Followerwonk: This tool helps you perform basic Twitter analytics, such as, who are your followers? Where are they located? When do they tweet? The tool can be accessed via http://followerwonk.com/

Esri: Esri's GIS (geographic information systems) is software to map, visualize, question, analyze, and interpret data to understand relationships, patterns, and trends (http://www.esri.com/).

Subject: Social Media Analytics

Module 4: Social Media Location & Search Engine Analytics

SEARCH ENGINES ANALYTICS

SEARCH ENGINES are the gateways to social media and help users search for and find information. To be more precise, a search engine is an Internet service or software designed to search information on the web that corresponds to a request (e.g., keywords) specified by the user. Considering that there are billions of websites over the web, search engines play a crucial role in helping us find the right information in a limited amount of time. Before shifting our focus to search engines analytics, let's understand different types of search engines.

TYPES OF SEARCH ENGINES

Based on the mechanisms they operate, search engine can be divided into three types: 1) Crawler-based, 2) Directories, and 3) Metasearch engines.

1. **CRAWLER-BASED**: As the name suggests, crawler-based search engines create their databases or lists automatically, without any human intervention. Examples of crawler-based search engine are Google.com and Bing.com. Crawler-based search engines are widely used to find and access content over the Internet. They operate in three steps: 1) web crawling, 2) indexing, and 3) searching.

Web crawling—Search engines start by collecting and storing information about web pages. This mechanism is termed web crawling. A web crawler (also known as web spider or bot) is a computer program or software specifically designed to collect and store data about websites for indexing.

Indexing—Indexing helps classify a website correctly for searching purposes. The data crawled or extracted is then indexed and stored in a database for quick access. Every search engine may follow different techniques for indexing web page data. Common indexing techniques include storing meta tags (which are used in the header of a web page and provide descriptions of the website) and keywords related to a website.

Searching—Searching is the final step in search engine operations. When a user requests specific information by entering keywords in a search engine, the search engine queries the index and provides a list of the most relevant web pages by matching it with the indexed keywords. However, it may not be that simple; search engines use a variety of factors to rank and provide a list of matching websites.

- A takeaway here is that in order to achieve good search results an organization must place keywords in section titles, images, and in the general content of its website. A keyword density of 5–8 percent (i.e., five to eight keywords per one hundred words) is an optimal number. Having important keywords embedded in a website enables a search engine robot to evaluate the website as being the most suitable site for the searched word.

However, if one repeatedly uses the same keywords or definitions in page content, it may be perceived by a robot to be spamming.

- Research has shown that the position of key words in a website, as well as their duplication, layout, and combination, impact web page visibility in a search engine, which can be improved by increasing the frequency of keywords in the title, the full text, and in both the title and full-text. In conjunction with key words, the overall design of a website is an important factor that must be taken into consideration when discussing search engine optimization.
- For example, flash animations, while aesthetically appealing, can negatively impact the SEO evaluation results because they cannot be indexed as easily by bots as more simply structured HTML content.
- For a corporation to better understand its Internet presence, its website statistics should also be checked on a regular basis so as to understand both how users access and utilize the site and also what impact site changes may have on these behaviors.

2. **DIRECTORIES**: The listings in directories are manually compiled and created by human editors. People who want to be listed in a directory submit an address, title, and brief description of their website, which is then reviewed by the editor and included in it. Some good examples of human-created directories are Yahoo Directory, Open Directory, and LookSmart.

3. **METASEARCH ENGINES**: Metasearch engines compiles and display results from other search engines. When a user enters a query, the metasearch engine submits the query to several individual search engines, and results returned from all the search engines are integrated, ranked, and displayed to the user. Examples of meta–search engines include Metacrawler, Mamma, and Dogpile. By integrating results from several search engines, metasearch engines are capable of handling large amounts of data and can help us save time by focusing on one search engine.

   Based on their scope, search engines can be divided into two types: 1) local and 2) global.

   1) **LOCAL SEARCH ENGINES**: A search engine is local in the sense that it is embedded within a website and only indexes and searches the content of that website. Amazon's Cloud Search or any other search engine embedded within a website is an example of local search engine.
   2) **GLOBAL SEARCH ENGINES**: Global search engines are used to search for content on the web. Google.com and Bing. com are examples of global search engines. However, note that global search engines can be localized. Google search, for example, can also embed within your website to help users find information on your website.

SEARCH ENGINE ANALYTICS

Generally, when we talk about search engine analytics, we mean two things:
1) search engine optimization and
2) search engine trend analysis.

SEARCH ENGINE OPTIMIZATION

- Search engine optimization (SEO) are the techniques used to improve a website's ranking in a search engine result page (SERP).
- A SERP is the list of the results returned by a search engine in response to a user's query.
- SERPs generally have two types of results: organic and nonorganic search results. Organic results appear mainly because of their relevance to the user's query. Nonorganic search results include paid advertisements.
- A study tested the effect of sponsored ad ranks on the click-through and conversion rates for an online retailer and found that top positions usually had higher click through rates, but not necessarily higher conversion rates.
- Social media marketers strive to develop search engine strategies to make their websites appear at the top of search results. It is important for their websites to appear at top (e.g., in the top ten) in the SERP, as users pay closer attention to the top results on search engines.
- The ranking becomes more crucial when the website is commercial by nature; that is, selling products or services. High rankings on SERPs can mean more Internet traffic to a website, which in some cases converts to more paying clients and higher return on investment.
- For social media marketers, it is important to understand the mechanism behind the SERP ranking. There may be variety of factors search engines take into account to rank websites, such as keywords and relevance. However, the most important factor that determines SERP ranking is the PageRank.
- PageRank is a mechanism (or an algorithm, to be more precise) used by Google search engines to rank websites' SERPs. The websites that rank higher are displayed on the top of the search results page.
- Google's PageRank algorithm predominantly relies on the quality of incoming hyperlinks (or in-links) to rank websites. A website, for example, with in-links from a famous website (e.g., cnn.com) will appear on the top of the SERP if compared with a website with no quality in-links or many low-quality in-links.
- To understand the in-link quality and number argument, consider Figure 12, where nodes represent web pages and lines represent in-links (arrowhead pointing to a page) and out-links (arrowhead pointing away from a page).
- The PageRank algorithm will place page B higher on the SERP, even though there are fewer in-links to B when compared to D. The reason for this ranking is that in-links to website B are from an important website; that is, A. Bottom line, your objective is to increase the number of qualities in-links to your website.

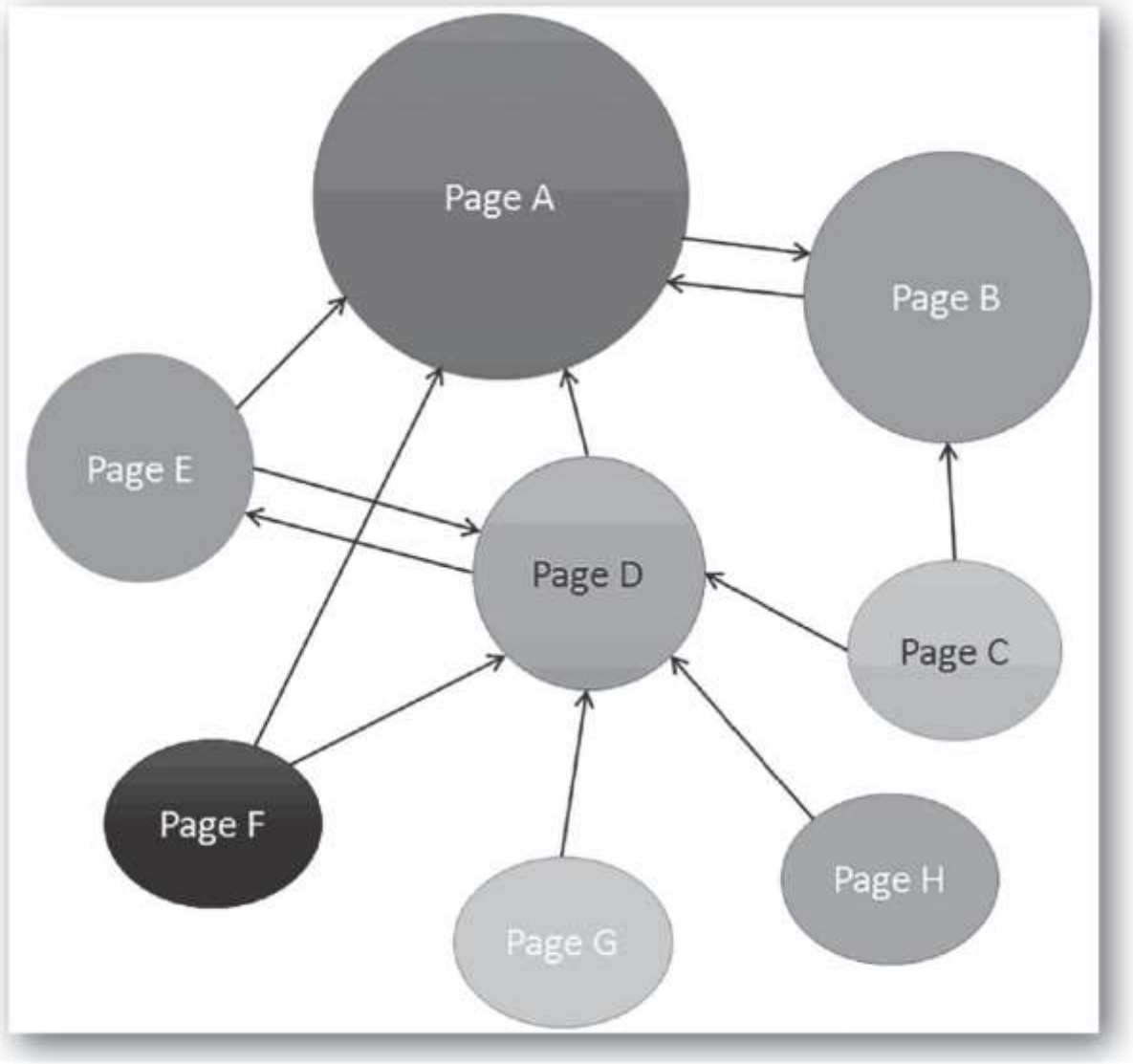

Figure 12. PageRank algorithm ranking example

By using, for example, Open SEO Stats (an extension for Google Chrome available at: http://pagerank.chromefans.org/), users can be determine the ranking of a website based on Google PageRank. Google PageRank uses a scale of 0 to 10, indicating the importance that the Google search engine allocates to the page. In addition to page ranking, Open SEO Stats also provides information about website traffic, hyperlink status, and speed of the page, among other things.

SEARCH TREND ANALYTICS

Search engine trends analytics deals with analysing and understanding the keywords people use in a search engine. Search engine data are gateways into the minds of customers. Through search engines, customers search for what they want, thus search trend analysis can provide value information to the social marketers.

When it comes to trends analytics, Google Trends (http://www.google.com/trends/) is one of the most convenient and comprehensive search engine trend analysis tools. Google trends use massive

amounts of search engine data to analyze the world's interests and predict trends. In the financial sector, Google Trends data, for example, has been used to detect "early warning signs" of stock market moves. In the health sector, Google Trends data has helped determine world flu epidemics. Engineers at Google.org, for instance, using Google Trends data, found a strong correlation among the searches for flu related topics and the numbers of actual flu cases circulating in different countries and regions around the world.

From a business perspective, Google Trends can help also answer a variety of questions, including the following:

- How people search for your brand?
- When does interest spike in your products or services?
- Which keywords drive more traffic?
- Which regions are interested in your brand?
- What are trending topics over the Internet?
- How are your competitors performing?

SEARCH ENGINE ANALYTICS TOOLS

Google Trends: Google Trends (http://trends.google.com/) is a search engine analytics tool.

Canopy: Canopy is multimedia analytics tool designed to support deep investigation of large multimedia collections, such as images, videos, and documents. More information on Canopy is available here: http://www.vacommunity.org/article32

Google Alerts: Google Alerts (https://www.google.com/alerts) is a content detection and notification service that automatically notifies users when new content over the Internet (e.g., social media, web, blogs, video and/or discussion groups) matches a set of search terms based on user queries. Users are alerted through e-mail. Find out about how to use Google Alerts.

Icerocket: Icerocket (http://www.icerocket.com/) specializes in blog searches and also captures activity on Facebook, Twitter, and Flickr.

Social Mention: Social Mention (http://socialmention.com/) is similar to Google Alerts, but it only focuses on social media sites, and you can choose to focus on particular areas, such as blogs. With Social Mention you can monitor for the appearance of particular keywords and it will give you information on related users, hashtags, and more.

TweetBeep: TweetBeep (http://tweetbeep.com/) is like Google Alerts for Twitter. Choose some keywords and receive daily search results via e-mail.

Subject: Social Media Analytics

Module 5: Social Information Filtering

How does a user process all the information? Some of it will be more useful than the rest, and the sheer volume often means a filter would be helpful to sort through it all. This chapter introduces several methods used to sort, filter, and aggregate information from social media using social connections.

Social Sharing and Social Filtering:

Social Sharing and Social Filtering use the interests of others, especially friends on social networks, to highlights information that is more likely to be of interest.

The reliance on large numbers of people to help complete a task is a type of crowdsourcing. From the "crowd" of people online, each contributes a tiny amount of work by sharing or voting on content, and the aggregate results are a valuable contribution.

Frictionless sharing refers to the transparent sharing of resources using social media services. The term became popularised following Mark Zuckerberg's announcement at the F8 developers conference in 2011 in which he described developments to Facebook that would allow "real-time serendipity in a friction-less experience". The term had previously been used in an application-independent way to describe "sharing that occurs without any additional effort required, for example, if a scholar is gathering resources for her own research, then using a social bookmarking tool is an effective tool for her as well as making the list public". This does raise privacy concerns for many users who do not want all their activity shared, but the practice is particularly popular among content publishers, like news sites, who see it as a new way to highlight interesting content and pull in readers. Aggregating behaviour that may show the "top-viewed" articles without information about who viewed them is one way around the privacy issues.

Automated recommender systems

Recommender systems are major parts of e-commerce sites and social media sites.

Even if the term recommender system is not a familiar one, nearly all Internet users will be familiar with them. These are the features of websites that suggest items a user might like. Amazon.com uses one to suggest other items a customer might want to buy. Netflix uses it to suggest movies that a subscriber might want to check out. Pandora uses it to automatically generate Internet music channels that match a user's taste. All of these personalized suggestions based on a user's previous activity come from recommender systems. They rely on explicit data, such as user ratings, or implicitly captured data from users' behavior such as making purchases or viewing an item.

Because good recommendations keep customers interested in a website and increase the likelihood that they will buy something, they have become big business. As more and more user-generated content comes online, new algorithms for generating recommendations are

created. These methods for generating suggestions are often quite complex, but the core idea behind many of them is to take data created by other people and personalize it for the individual user. It is an excellent example of aggregating information, especially ratings of items, in a social way.

Traditional recommender systems

Recommender systems basically work in one of two ways: suggesting items similar to the ones a person likes or suggesting items liked by people who are similar to the user. They might look at all the items that a user has rated and then look for items that are similar to the things the user likes. This is how Pandora, the online music streaming service, works. A user starts with a song or artist, and Pandora creates a musical profile of it. Then, Pandora selects songs that are similar in profile and plays those. If the user gives a new song a thumbs up, the profile of that song is combined with the existing profile to create a new set of attributes that the user likes. If the user gives a song a thumbs down, then the attributes of that song are deemphasized in the profile. This tactic of finding items similar to what the user is known to like is called item-based or model-based recommendation.

Item-based recommendation is not very social in that it does not rely on other people very much, but the second type of recommender systems relies entirely on other people's actions. These work by finding people who have similar tastes to the user and then recommending items that those people like. This is called collaborative filtering. At its core, collaborative filtering looks at each pair of users, finds the items that both people have rated, and computes a similarity score for the two people based on their ratings. That similarity measure is then used to give similar people more say in how much the user might like a new item. Consider this simple example of collaborative filtering. A user, Alice, has rated a set of movies. Two other users, Bob and Chuck, have also rated those movies. These are shown in Table 13.1.

Table 13.1 Ratings by Three Users for Five Different Movies. Ratings are on a 1 5 Scale

	Star Wars	Jaws	Wizard of Oz	The Godfather	2001
Alice	5	4	3	3	1
Bob	3	5	2	5	1
Chuck	4	3	2	2	2

There are many ways to compute similarity with Alice. One option would be the average difference between ratings of these movies. In this example, Bob has an average difference of 1.2 with Alice, while Chuck has an average difference of 1.0. A more common measure of similarity is the correlation between the ratings. Chuck's ratings are 1 point lower than Alice's for every movie except 2001, where it is 1 higher. His ratings track very closely with hers. Bob, on the other hand, does not seem to follow any pattern of being higher or lower with respect to Alice. This idea is captured by the Pearson Correlation Coefficient, a simple statistic that measures how well aligned two sets of values are. You can compute the Person correlation in

most standard spreadsheet applications, including Microsoft Excel. It is always a number between 1 and 1, where a higher positive number indicates a high similarity and a negative number indicates preferences that vary in the opposite direction. In this example, the correlation between Alice and Bob is 0.26 and the correlation between Alice and Chuck is 0.83. Because correlation is commonly used in collaborative filtering, the rest of this example will use those values. Now assume Alice wants to know how much she might like the movie Vertigo, which she has never seen. Both Bob and Chuck have seen it. Bob rated it a 3 and Chuck rated it a 5. What would be a good recommendation to Alice for how much she will like it? One option is to show the average rating for the movie, which is a 4 in this case. However, that does not take into account that Chuck is more similar to Alice than Bob is. A simple example of collaborative filtering will use the correlation numbers to compute a weighted average. Bob and Chuck's ratings will be multiplied by their correlation with Alice, and that total will be divided by the sum of the weights.

$$\frac{\begin{matrix}\textit{Bob} & \textit{Chuck}\\ 0.26*3 + & 0.83*5\end{matrix}}{0.26 + 0.83} = 4.5$$

Notice that this weighted average comes out higher than the simple average. That is because Chuck gets more weight, and since he is more similar to Alice, his higher rating of the movie is given more consideration. Thus, the recommended rating of Vertigo for Alice is 4.5 stars.

The ratings produced by recommender systems can be used directly to indicate to a user how much they might like a particular item, or they can be used to sort items, showing those that seem most promising higher up in a list. They can also be used to filter out items a user is unlikely to like.

Social recommender systems

Collaborative filtering is an early example of how algorithms can leverage data from the crowd. Information from a lot of people online is collected and used to generate personalized suggestions for any user. These techniques were originally developed in the 1990s and early 2000s. Since the availability of this data has increased with the rise of social media, recommender systems have started to consider social connections in addition to similarity.

Simple examples of social recommendations can be found on many social networking websites. For example, on Twitter, when a user searches for a term, the search results can be shown in three ways: all tweets that match the search, "top" tweets, as determined by Twitter, or tweets only posted by people the user knows. This simple social filter excludes anything from unknown people, since it may be of less interest.

Friend recommenders are also common in social networking websites. Facebook prominently features a "People You May Know" section, which is essentially a recommendation of people

to add as friends. This uses social network data to guess at what edges might be missing from a network. For example, if you are friends with 9 out of 10 densely connected people, it is likely that you are also friends with the 10th person.

Social relationships can also be used with collaborative filtering algorithms. The similarity measure that these algorithms traditionally use can be replaced with a variety of statistics taken from the network. Using trust or tie strength would give more weight to people who are close to the user and likely share similar opinions. Trust has been particularly well studied for making recommendations, and systems exist that leverage it for applications as diverse as recommending a movie to recommending mountain ski routes.

UNDERSTANDING SOCIAL MEDIA AND BUSINESS ALIGNMENT

As with any other technology, aligning social media objectives and goals with the objectives of the organization should be the starting point of any social media analytics initiative. The alignment of social media analytics with business objectives can be seen as analogous to the famous Chinese Yin and Yang philosophy, where two seemingly opposing forces (in this case, social media and business) complement and reinforce each other (Figure 13).

Figure 13. Aligning Social Media Analytics with business goals (Yin and Yang philosophy)

Figure 14 provides example scenarios for aligning social media with business objectives. If the business goal is to understand customer sentiments expressed over social media, the social media analytics should be designed to facilitate this objective. It may require, for example, tools and skills for extracting and analyzing tweets or comments posted on a Facebook fan page. Or, for example, if your business objective is to identify influential social media customers and their position in the network, your focus should be on social media networks.

Example Business Question	Layers of Interest	Data Source	Example of Tools
Is the social media conversation about our company, product, or service positive, negative, or neutral?	Text Analytics	Tweets Comments Blog posts Reviews	Discovertext Lexalytics Semanteria
Which content posted over social media is resonating more with my customers?	Actions Analytics	Likes Shares Views Mentions	Google Analytics Hootsuite
Who are our influential social media nodes, and what is their position in the network?	Network Analytics	Fan network Follower network	NodeXL Flocker Netlytic Mentionmapp
How is our mobile app performing?	Mobile Analytics	Total sessions New users Time spent	Countly Mixpanel Google Mobile Analytics
Where are our social media customers located?	Location Analytics	Geo-map IP address GPS	Google Fusion Table Tweepsmap Followerwonk
Which social media platforms are driving most traffic to our corporate website?	Hyperlink Analytics	Hyperlinks In-links Co-links	Webometrics Analyst VOSON
Which keywords and terms are trending?	Research Engine Analytics	Trending topics	Google Trends

Figure 14. Aligning analytics with business objectives

Following are the areas for Understanding social media and Business Alignment:

- **Marketing and advertising**: Businesses can use social media to reach potential customers and promote their products or services. This can include creating a social media presence, posting content, and running ads on social media platforms.
- **Customer service**: Businesses can use social media to provide customer support and answer questions or resolve issues. This can include monitoring social media for mentions of the business and responding to customer inquiries or complaints.
- **Public relations**: Businesses can use social media to manage their reputation and communicate with the public. This can include responding to negative reviews or feedback, as well as sharing news and updates about the business.
- **Talent recruitment**: Businesses can use social media to find and attract potential employees by posting job openings and engaging with potential candidates.

- **Lead generation**: Businesses can use social media to generate leads and connect with potential customers by sharing valuable content and offering incentives for users to provide their contact information.
- **Market research:** Businesses can use social media to gather insights about their customers, competitors, and industry trends by monitoring social media conversations and conducting surveys.
- **Product development:** Businesses can use social media to gather feedback and ideas for new products or features by engaging with customers and asking for their input.

SOCIAL MEDIA ANALYTICS ALIGNMENT MATRIX

The extent and breadth of your social media analytics alignment with business goals will be determined by a variety of factors, including the availability of technical, financial, administrative, and leadership resources, and its potential to achieve business goals.

Here we use a simplified social media analytics alignment matrix provided in Figure 15. On the Y axis of the matrix is "resource availability," which refers to the availability of financial, technical, administrative, and leadership resources for social media analytics. On the X axis of the matrix is the impact of social media analytics alignment in terms of its potential to achieve business goals (or its potential to generate economic value and return on investment). Depending on the two variables (i.e., resources availability and its potential), your social media analytics alignment with business goals can fall into four possible quadrants. Your alignment resides in the "highly aligned" quadrant, for example, when leadership, financial, administrative, and technical resources are available to leverage and (sustain) social media analytics and its potential is high in terms of achieving business goals. For instance, mining the seven layers of social media data is technically and financially demanding, but rewarding in terms of the creation of economic value to the firm. And your social media analytics alignment efforts reside in the "not aligned" quadrant when its potential to achieve business goals and your resource availability is low.

Figure 15. Social Media Alignment Matrix

The social media analytics alignment matrix will guide us throughout the social media strategy formulation process. The alignment matrix is flexible. One can replace the variables at both the axes with any other variables of interest. For example, we can place criticality of social media analytics (the extent to which the analytics is critical to the business) on the Y axis and sensitivity of the analytics (e.g., in terms of security, privacy, or ethics) on X axis and determine the extent of your alignment. Your social media analytics alignment, for instance, will be considered "highly aligned" if it is business critical, but less sensitive.

ROLE OF CIO AND IT MANAGEMENT

Senior IT executives, particularly the CIO, play an important role in envisioning and creating aligned social media analytics strategy. The CIO is the person in charge of managing and aligning information communications technologies (ICTs) to achieve business-wide goals. The role of CIO has evolved from a technical guru to an informed leader, communicator, and strategic thinker. For a sustained strategic IT–business goals alignment, a CIO should possess the following skills and competences:

Strategic Thinking and Evaluation

✓ Business and policy reasoning

✓ IT investment for value creation

✓ Performance assessment

✓ Evaluation and adjustment

Systems Orientation

✓ Environmental awareness
✓ System and social dynamics
✓ Stakeholders and users
✓ Business processes
✓ Information flow and workflow

Appreciation for Complexity

✓ Communication
✓ Negotiation
✓ Cross-boundary relationships
✓ Risk assessment and management
✓ Problem solving

Information Stewardship

✓ Information policies
✓ Data management
✓ Data quality
✓ Information sharing and integration
✓ Records management ✓ Information preservation

Technical Leadership

✓ Communication and education
✓ Architecture
✓ Infrastructure
✓ Information and systems security
✓ Support and services
✓ IT workforce investments

Measure of Success for a Company's Social Media Campaign:

- **Counts**: This includes the number of fans, followers, or friends, as well as the number of views, likes, or similar indications of favorable opinions on the company's social media content.
- **Social sharing**: This includes the number of times the company's content is shared, mentioned, or retweeted on social media platforms.

- **Engagement rate**: This is the number of engagement activities (likes, shares, etc.) divided by the number of friends, followers, or fans, and indicates the level of engagement of the company's social media audience.
- **Interaction:** This includes the number of customers with whom the company has engaged, the number of conversations, and the length and resolution of those conversations.
- **Referral rates:** This is the amount of traffic driven to the company's website from its social media presence, as measured through click-throughs or website analytics.
- **Importance and influence of users:** This includes metrics such as centrality or the number of friends, which indicate the influence and importance of users in the company's social media network.

Social Media KPI

- **Reach:** The number of people who see a business's social media content, including followers, friends, and other users who come across the content.
- **Engagement:** The level of interaction with a business's social media content, including likes, comments, shares, and other actions taken by users.
- **Traffic:** The number of users who click on links from a business's social media content and visit its website.
- **Conversion rate:** The percentage of users who take a desired action after visiting a business's website from its social media content, such as making a purchase or signing up for a newsletter.
- **Customer satisfaction**: The level of satisfaction of a business's customers, as measured through social media interactions or surveys.
- **Lead generation**: The number of leads generated through social media, such as users who provide their contact information in exchange for an offer or resource.
- **Return on investment (ROI)**: The financial return on a business's social media efforts, calculated as the profit gained divided by the cost of the social media campaign.
- **Cost per acquisition (CPA):** The cost of acquiring a new customer through social media, calculated as the cost of the social media campaign divided by the number of new customers acquired.

FORMULATING A SOCIAL MEDIA STRATEGY

Formulating a social media strategy is not much different than overall information technology (IT) strategy of an organization. The purpose of formulating social media strategy is to create rules and procedures to align your social media engagement with business goals. Planning an aligned social media strategy should follow a strategy formulation process similar to that used by IT management as suggested by Luftman et al. (2004), though some additional steps are needed to account for the unique nature of social media technologies.

STEPS IN FORMULATING A SOCIAL MEDIA STRATEGY

The following steps will lead to the formulation of a sound social media strategy.

1. **Define objectives:** Identify the specific goals that the business wants to achieve through social media, such as increasing brand awareness, generating leads, or improving customer satisfaction.
2. **Identify target audience:** Determine the demographics and interests of the business's target audience, including age, gender, location, and social media habits.
3. **Research competition:** Analyze the social media presence and strategies of the business's competitors to understand what is and is not working in the industry.
4. **Choose social media platforms:** Select the social media platforms that are most relevant to the business's target audience and objectives.
5. **Create a content calendar**: Plan and schedule the types of content that the business will post on social media, including text, images, videos, and links.
6. **Engage with followers**: Monitor and respond to comments and inquiries from followers, and encourage user-generated content and interactions.
7. **Analyze and adjust**: Use social media analytics tools to track the performance of the business's social media efforts and make adjustments as needed to improve results.

MANAGING SOCIAL MEDIA RISKS

Engaging through social media introduces new challenges related to privacy, security, data management, accessibility, social inclusion, governance, and other information security issues. Risk, in simple words, is the possibility of losing something of value such as, intellectual or physical capital. A comprehensive definition of risk is provided by National Institute of Standards and Technology (NIST), which states that "risk is a function of the likelihood of a given threat source's exercising a particular potential vulnerability, and the resulting impact of that adverse event on the organization." Here we will focus on the risk arising from social media use and define it as the potential of losing something of value (such as information, reputation, or goodwill) due to the use of social media tools and technologies.

Social media–related risks needs to be managed properly, both from the strategic and technological points of view. To minimize the damage, organizations need proactive, rather than reactive, social media risk-management strategy. Following are the steps for managing social media risks:

- **Develop a social media policy**: Create guidelines for employees on how to use social media in a professional manner and how to handle sensitive or confidential information.
- **Monitor social media activity**: Regularly review the business's social media presence and activity to identify any potential risks or issues.
- **Respond to negative feedback**: Address any negative comments or reviews in a professional and timely manner, and work to resolve any issues that may arise.
- **Protect personal information**: Ensure that personal information, such as customer data, is kept secure and not shared on social media without proper consent.

- **Stay up to date with legal requirements**: Understand and comply with relevant laws and regulations, such as those related to privacy, advertising, and consumer protection.
- **Train employees**: Educate employees on how to handle social media risks and best practices for using social media in a professional manner.
- **Have a plan in place:** Have a plan in place for handling social media crises, including identifying a team to manage the situation and establishing protocols for communication.

Social Media Risk Management Framework

A simple but effective way to proactively manage social media risks is through the social media crisis management loop (Figure 15), which includes four iterative steps: 1) identify, 2) access, 3) mitigate, and 4) evaluate. In the identification stage, potential risks are identified, which, in the assessment stage, are assessed and prioritized in terms of probability of occurrence and impact on the agency. In the mitigation stage, risk mitigation strategies are formulated and implemented. Finally, periodic assessment and reviews are carried out in the evaluation stage of the risk-management loop. Below, we discuss each step in detail.

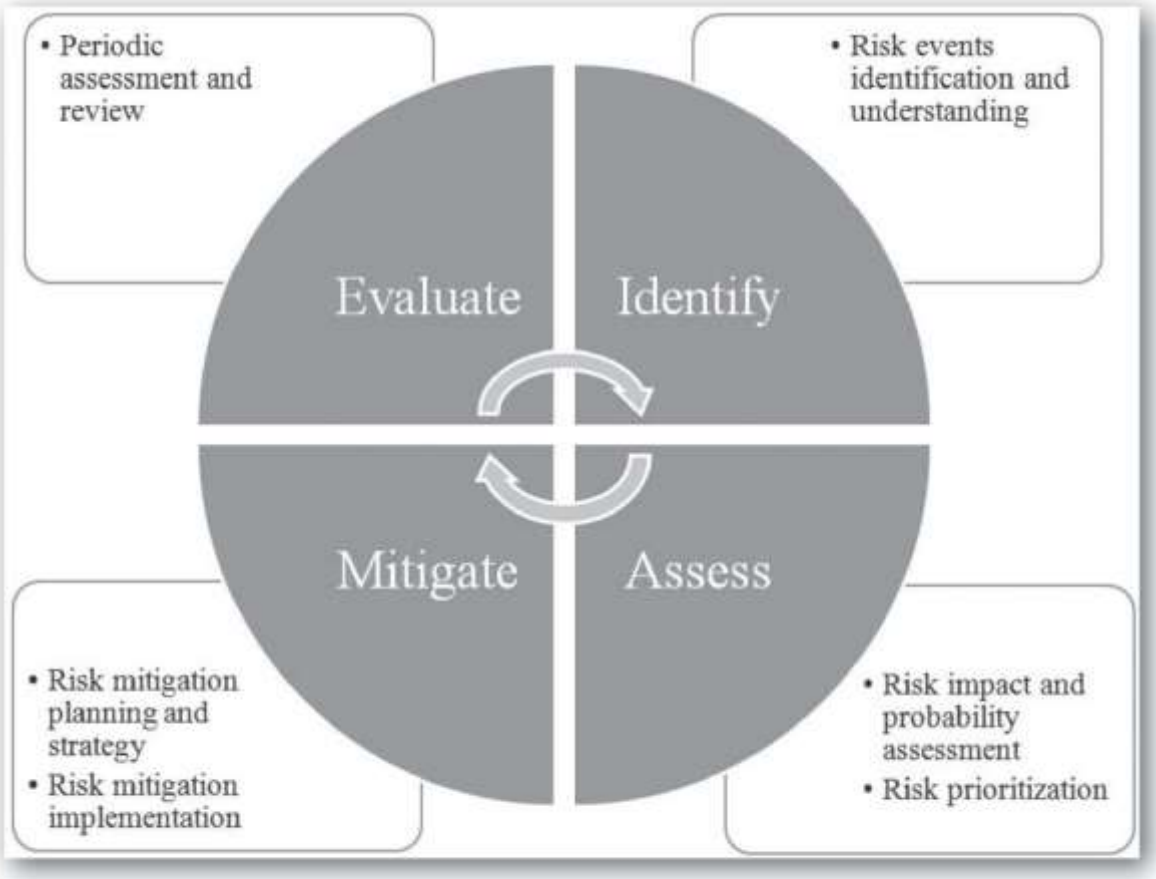

Figure 16. Social Media Risk-management Framework

RISK IDENTIFICATION

Risk identification is the process of identifying social media threats in terms of vulnerabilities and exploits that could potentially inhibit your organization from achieving its objectives. At this stage, your goal is to identify potential accidental or malicious risks that can come from within or outside the company. Examples of social media–related security breaches are hacking, information leaks, phishing, and impersonation. Phishing and hacking are examples of malicious outsider attacks. Famous social media platforms (such as Facebook, Twitter, and YouTube) are riskier than less famous ones (Webber 2012). According a study based on surveys and interviews with ninety-nine professionals and thirty-six companies (Webber 2012), the main social media risks identified were
1) damage to reputation,
2) release of confidential information,
3) legal, regulatory, and compliance violations,
4) identity theft and hijacking, and 5) loss of intellectual property.
Other potential social media risks include malware, loss of privacy, and social engineering attacks.

RISK ASSESSMENT

Risk assessment is "the process of assessing the probabilities and consequences of risk events if they are realized" (MITRE 2014). The risk assessment process determines the likelihood of a social media risk event that could impact the organization economically, technically, politically, and socially. The potential risks identified in the earlier step are priorities and ranked based on probability of occurrence and impact on an organization (Garvey 2008).
Probability (P) is the likelihood of occurrence of a risk event and can take a value from 0 to 1. Probability can, for example, be assigned to risks events as follows.

Certain to occur (P=1)—The risks with a P value equal to 1 are the risks that will certainly happen. In other words, they have a 100 percent chance of occurring.

Extremely sure to occur (P=> 95 <1)—The risks, for example, with a probability value greater than 0.95 and less than 1.0 can be considered as "extremely sure to happen" risks. In other words, they have a 95–100 percent chance of occurring.

Almost sure to occur (P= > 0.85 <= 0.95)—The risks with a probability value greater than 0.85 and less than or equal to 0.95 can be considered as "very likely to occur" risks. The can be said to have an 85–95 percent chance of occurring.

Very likely to occur (P=> 0.75 <=0.85)—These are the risks with a 75–85 percent chance of occurring.

Likely to occur (P=> 0.65 <=0.75)—these are the risks with a 65 to 75 percent chance of occurring.

Slightly likely to occur (P=> 0.55 <=0.65)—these are the risks with a 65 to 75 percent chance of occurring.

Evenly likely to occur (P=> 0.45 <=0.55)—these are the risks with a 45 to 55 percent chance of occurring.

Risk Impact of a risk event can be characterized as 1) severe, 2) significant, 3) moderate, 4) minor, or 5) minimal. A risk event is considered severe, for example, if it has devastating economical, technological, political, or social impact on your agency. And a risk is considered minimal if its impact is very low or negligible.
Based on the impact and probability, social media risks can be prioritized as 1) high, 2) medium, or 3) low .

High priority risks—The risks that, if they happen, will have severe economic, technological, political, or social impact on your agency. These are the risks that needs immediate attention and should be managed carefully.

Medium priority risks—The medium-probability risks that, if they happen, will have considerable economic, technological, political, or social impact on your agency.

Low priority risks—The low probability risks that, if they happen, will have low economic, technological, political, or social impact on your agency.

As shown in the Figure 17, you can assign other probabilities in a similar way.

Probability (P)	Chance of occurrence	Priority
P=1	Certain to occur	High Priority Risks
P=> 0.95 <1	Extremely sure to occur	High Priority Risks
P=> 0.85 <= 0.95	Almost sure to occur	High Priority Risks
P=> 0.75 <=0.85	Very likely to occur	High Priority Risks
P=> 0.65 <=0.75	Likely to occur	High Priority Risks
P=> 0.55 <=0.65	Slightly likely to occur	Medium Priority Risks
P=> 0.45 <=0.55	Evenly likely to occur	Medium Priority Risks
P=> 0.35 <=0.45	Less than an even chance	Medium Priority Risks
P=> 0.25 <=0.35	Less likely to occur	Low Priority Risks
P=> 0.15 <=0.25	Not likely to occur	Low Priority Risks
P=> 0.00 <=0.15	Certainly sure not to occur	Low Priority Risks

Figure 17. Risk probability and prioritization assessment

RISK MITIGATION

The risks prioritized and ranked in the earlier stage should be physically, technically, and procedurally managed, eliminated, or reduced to an acceptable level. Dependent on the nature of the risks, different strategies should be used; for example, accidental risks posed by employees (e.g., posting copyright material online or tweeting some confidential information) can be eliminated by training, awareness programs, and by having sound social media policy in place. Hacking attacks, for example, can be mitigated using updated antivirus systems and by creating an extra layer of security, such as a two-mode authentication technique. Typical risks mitigation strategies are:

1.Risks management governance—New governance structures, roles, and policies should be created within your business for properly managing social media risks. These activities may involve identifying and empowering a social media risk-management manager, developing a business-wide risk-management strategy, identifying actions and steps needed to implement the strategy, and determining the resources required to mitigate the risks (Garvey 2008). Create a social media risk-management governance structure by involving all key departments, including IT, finance, public relations, human resources, legal, and communications. All of these components play an important role in identifying and mitigating social media risks.
2.Training and awareness—Provide education and spreading awareness on legal issues such as copyright, intellectual property, defamation, slander, and anti-trust issues.
3.Social media policy—Create a sound social media policy that outlines the relative rights and responsibilities of employers and employees.

4.Secure your social media platforms—Secure your social media platforms to minimize the impact or likelihood of the risk.
The following are some techniques you can use to secure your social media platforms.

Use strong passwords—to protect your social media accounts (Twitter, Facebook, YouTube, blogs, etc.) always use strong passwords. A password is considered strong when it:
✓ Is at least ten characters long.
✓ Has a combination of uppercase and lowercase letters, numbers, and symbols.
✓ Does not include your personal information such as phone numbers, birthdays, name, etc.
✓ Does not use common words such as "mypassword," "ilikeyou," etc.
✓ Does not use alphabetical sequences (such as "abcd1234") or keyboard sequences (such as "qwerty")
✓ Is not reused across websites; that is, your Twitter account password should be unique to Twitter. ✓ Is memorized or kept in a safe place if written.

Two-mode authentication—It provides an extra layer of security that uses your phone to protect your account. For example, if your account is compromised or someone figures out your password, they will still not be able to access to your account unless they have physical access to your phone. Each time you login in from an unknown browser or computer, you will need to provide a security code to access your account (unless you list the device as secure).

Trusted contact—The trusted contact is an account recovery feature provided by Facebook to help you access your account securely through your friends if you have trouble accessing your account.

Review your login history—It is a good practice to regularly review your account login history and location.

Login notification—enable your Facebook login notification so that you can be notified through e-mail or text message when your account is accessed.

Disable or revoke third-party apps—Your information (e.g., friends list and profile) is available to third-party applications (or apps for short) running over Facebook.

RISK EVALUATION

Social media risk management is a continuous process. In the face of rapid technological, political, and social change, social media risks should be periodically reviewed. Your risk-management strategy, procedures, and techniques should be continuously updated in response to the emergence of new social media platforms, social changes, and potential new risks. The

continuous evaluation and monitoring effort will make sure that the initial assumption made about the external and internal risks are still relevant.

* Social Media in the Public Sector

Traditional media is broadcast, messages go out, but there is little interaction with the audience. Information cannot easily be personalized, and quick, real-time responsiveness to audiences is difficult or impossible in most traditional media.

Social media, on the other hand, is based around interaction. It allows personal communication, personalization of broadcast messages, and it can be updated quickly to keep people apprised of a rapidly changing situation.

One could say that traditional media is one-to-many communication, while social media is many-to-many. These benefits are making social media increasingly popular as a communication mechanism in the public sector among government, politicians, emergency-response officials, and others.

* Analyzing public-sector social media.

When analyzing public sector use of social media, it can be approached as an analysis of

1) how public-sector users are taking advantage of the technology or
2) how people are talking about public-sector related topics.

* Analyzing individual users

Public-sector users may be individuals, such as elected officials, local organizations, such as schools or libraries, or government agencies.

A social media user, whether an individual or organization, chooses to use social media for many reasons, but these are three major types of use:

1) Broadcast / Sending information :- A user may want to share information with followers or friends. These may be updates, requests, or other information. This use leverages social media because the audience is presumably interested in what the user has to say. These broadcast messages may be used to simply send information to followers or to engage in one of the following interaction types.

2) Request Feedback / Input :- Social media audiences may be willing to share information, whether it is an opinion about an issue, information they have about a crime, or posting the location of an event. A public-sector users may want to gather this information from its audience, and thus may request the feedback or input in their posts. They may also utilize social media as an open channel for people to send them comments.

3) Conversation Interaction :- Unlike a request for feedback, this type of use encourages conversation or interaction directly between the user and individual members of the social media audience. This may be an elected official having a conversation with constituents or a librarian talking to a patron through social media to help recommend books. Unlike requests for feedback or input, where an audience member's input is simply accepted and processed, this type of interaction supports back-and-forth communication between the user and audience member.

There are many ways to analyze how an individual organization is using social media. Based on your particular interest, you may create your own hypotheses and questions, but the following are general guiding questions.

- Who is doing the posting?
- Who are the target audience members?
- Why is the audience engaged in social media with the organization? What type of content or interaction is the audience interested in?
- What are the goals of the user? Which of the three interaction methods above are they using?
- How is the user using social media?
- Do the user's actions support the goals?

Subject: Social Media Analytics

Module 6: Social Media Analytics Applications and Privacy

Case study: Social media to solve an attempted child abduction

Just before 4 P.M. on July 17 2012, a 10-year-old girl and her 2-year-old brother were walking home from buying flavoured ice in their South Philadelphia neighbourhood. A man in a white car followed them for several blocks before parking and approaching the pair from behind. He grabbed the girl, covering her mouth and carrying away from her brother. The girl fought back, kicking, and biting her attacker, and broke free of the man's grip. He dropped her and then quickly fled the scene.

The children did not know the abductor, but the incident was captured on several surveillance videos. The Philadelphia Police Department's Special Victims Unit released the videos to the public less than a day after the event occurred. Figure 14.1 shows a still from one of those videos. The police immediately started receiving tips via social media channels. Within hours of posting the video, the man turned himself in, claiming that he "felt that he could not walk, talk or breathe out there," according to Philadelphia Police.

FIGURE 14.1

A scene from the surveillance video released by the Philadelphia Police Department on YouTube and through other social media to help capture the man who attempted to abduct a 10-year-old girl.

The Philadelphia Police Department has been a pioneering user of social media, actively using YouTube, Facebook, and Twitter to gather information about crimes. They also have smart phone apps that let people report incidents and find local police stations. At the time of this abduction, the department reported catching 87 suspects through social media usage. In February 2012, they caught the abductor and rapist of a 6-year-old girl within 16 minutes of posting the suspect's photo. In another case, a suspected murderer was turned in by his mother after she saw his photo.

To analyse the Philadelphia Police's use of social media, the guiding analysis questions can provide help. In terms of use, they are taking advantage of social media both to broadcast to a large audience

and to receive input from their audience. They are generally not having conversations in the social media, since the nature of their work means it is often better for a police officer to contact a person directly and privately to have an extended interaction.

Looking at the YouTube, Twitter, and Facebook accounts of the Philadelphia Police, we see that they are generally posting crime alerts and videos, to redirect other social media users to 911 dispatches if they are having an emergency, and to occasionally share departmental news. To be effective, they must not share too much information such that audience members would be overwhelmed, or information that is not relevant to them. The Philadelphia Police post to twitter usually less than 10 times a day, ensuring that they do not overwhelm users with too much content. But is social media also an effective way to receive information from people?

Another case from Philadelphia holds some clues. A man was attacked on a city bus, but no one on the bus was willing to assist him during the attack or call 911. However, after police posted the video of the attack online, several witnesses identified the suspect. They were more willing to report the event electronically than they were to get involved at the time.

A full analysis would look at a more fine-grained breakdown of the types of content being posted and the frequency. It would also look at the people who follow the Philadelphia Police. Demographic information, like location, may suggest some reasons they are interested in the content. Surveying people about why they follow the department may also provide more insight into how the Philadelphia Police are effective, what they might do differently or better, and how other departments might interact online to be similarly effective.

By understanding the attributes of the social media users and the interactions between police and users, the police can optimize their social media strategy. Knowing the kind of people who are involved in an organization's social media and knowing their interests allows an organization to post content that keeps the audience engaged, that takes advantage of their knowledge, and that thus allows the organization to more effectively get its message out.

The Philadelphia Police is only one example of how police and emergency response officials may use social media. Social media also allows people to report information about emergencies on location. As discussed in Chapter 12, Location-Based Social Interaction, mobile social applications are allowing people to report locations of wild fires. Systems have been proposed that leverage social media in communities to share information about ongoing incidents (Wu et al., 2008).

Case study: Congressional use of twitter

In mid-2012, nearly 400 members of the U.S. Congress had Twitter accounts. Although all these congresspeople represent constituents and have similar duties and goals in their positions, they use Twitter in very different ways. Some have staff members post to their accounts, generally sharing links to press releases and other official information. Other congresspeople tend to their accounts themselves, carrying on lively conversation with their constituents about issues, bills under consideration, and current events. While analysis of each account will reveal these differences in usage, analyzing all members of Congress together paints an interesting picture of the type of content being shared by these

similar users. This case study is a summary of work first presented in (Golbeck, Grimes, and Rogers, 2010).

Many insights are available by analyzing the way a specific group uses social media. The first step is to understand the needs of the group. What type of information do they want to convey? What type of interaction do they want to have with users? What type of activities do they undertake offline that might be well supported by social media? Then, analyzing how they are actually using social media may reveal a number of insights. Their existing utilization can be compared to their needs. Are they using it in ways that meet those needs? Are there needs that are not being met? Is social media being used to create new types of interaction? Could social media be leveraged to accomplish tasks for which it is not being leveraged?

Researchers collected over 6,000 tweets posted by members of the U.S. Congress over a six-month period. They read each tweet and categorized it into one or more of the following categories:

- Direct Communication a message directed at a specific person either with the @id convention or in the text of the message. Direct Communication was divided into two mutually exclusive subclasses.
  - Internal Communication This included messages from one congressperson to another or from a congressperson to a staff member.
  - External Communication All other messages, such as those to constituents, were marked as external communication.

• Personal Message These are non-business-oriented messages or notes, such as holiday greetings or other personal sentiments.

• Activities A message reporting on the congressperson's activities was divided into two mutually exclusive subclasses.

- Official Business: This included any official business in Congress, including voting, committee meetings, or making speeches on the house floor.
  Example Tweet: keithellison: Marking up the Credit Cardholder's Bill of Rights in the Financial Services Committee
- Location or Activity: This code was used when a Congressperson was describing non-official activities including trips, meetings with constituents, lobbyists, or non-Congressional organizations, or activities in the home district.
  Example Tweet: neilabercrombie: @neilabercrombie just completed weightlifting workout at the Nuuanu Y. Advertiser featuring him on July 10; it's part of a regular feature.

• Information This code describes a message that provides a fact, opinion, link to an article, position on an issue, or resource.

Example Tweet: greshambarrett: Barrett announces 10 campaign events: Congressman Barrett will campaign in his district this week, http://tinyurl. com/6puxze.

• Requesting Action When a congressperson requests constituents to take some action like signing a petition or voting, the message is coded this way.

• Fundraising Messages occasionally ask for donations and contributions, and we code those as fundraising.

• Unknown Some messages cannot be classified, as when they are only URLs with no text, test messages, or other mistakes such as a single character.

This is an example of how a finer-grained content analysis works. Researchers develop these types of categories by going over their data several times (in this case, reading the tweets) and coming up with a coherent and comprehensive set of categories to describe the content. Then, the data is reviewed again, placing each item in the appropriate category. This process is called coding. Ideally, two or more independent researchers code the data. Then, their agreement is calculated (called inter-rater reliability). A high inter-rater reliability indicates an accurate coding of the data.

After each tweet was categorized, patterns emerged about how these congressional users were taking advantage of Twitter. The vast majority of posts were Information (sharing links, opinions, or facts) or Locations and Activities (posts about unofficial activities the congressperson was doing).Perhaps surprisingly, the members of Congress did almost no political fundraising over Twitter, despite the fact that the collected tweets covered an election cycle. Figure 14.2 shows the breakdown of tweet types that the researchers found.

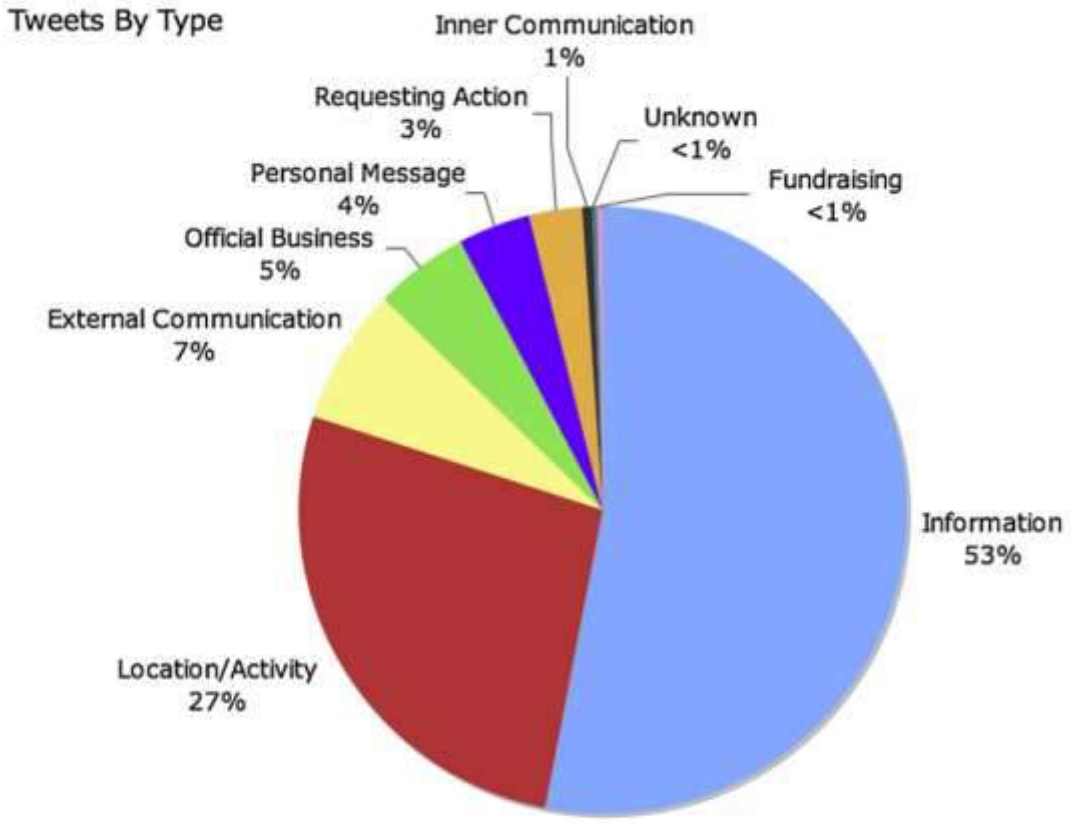

FIGURE 14.2

Types of tweets posted by members of the U.S. Congress as found in Golbeck, Grimes, and Rogers (2010).

Of course, as social media becomes more popular and widely understood, and as politicians and campaigns become more savvy with the technology, the patterns of use are likely to change. This case study serves as an example of how analysis of a group of users can be performed; it illustrates the type of activities common to the group, even when there are differences between individuals.

Case study: Predicting elections and astroturfing

The examples above have analyzed individual accounts or groups of individual accounts. This case study flips the perspective, analyzing behavior from a large and diverse set of social media users to understand public opinions about public issues.

While the study of how elected members of the U.S. Congress used Twitter showed little effort to raise campaign funds, social media has become increasingly important for political campaigns. The 2008 election of Barak Obama is largely cited as the first time a campaign truly embraced and used social media to reach voters. The power of social media to allow personal access to interested voters is powerful for campaigns, and the trends of how users discuss a campaign can provide valuable insight into what the public is thinking about an issue or an election. As a result, both individuals and the media have started using trends on social media as a way of understanding public opinion.

This has led some people to think that elections can be predicted based on how often a candidate is mentioned in social media and how positively or negatively that candidate is discussed. On the surface, this seems to be a valid approach. If social media is full of positive comments about Candidate X and there are fewer good posts about Candidate Y, then it would appear that Candidate X has more support and thus is more likely to win an election. There was also some anecdotal support for this technique, including the 2009 German elections (Tumasjan et al., 2010) and in the 2010 U.S. congressional elections (Livne et al., 2010).

However, further studies showed that using social media had only a slightly better-than-chance success rate at predicting elections (Metaxas, Mustafaraj, and Gayo-Avello, 2011). Furthermore, the volume of social media posts about a candidate were not necessarily representative of the public's opinion or conversation overall. A vocal minority could often overwhelm a silent majority, as was observed on Twitter in a 2010 special election for a Massachusetts Senate seat (Mustafaraj et al., 2011).

Although the trends and popular topics on social media may not reflect the public's opinion in general, it is still powerful to see a lot of discussion about an issue, particularly if it is favoring one side. Grassroots efforts often utilize social media to build interest in their causes and rally support. Since social media is inexpensive to use and can reach a large audience, it can be a very effective tool for gathering support and drawing attention to an issue. The success of grassroots movements online and the attention people are willing to pay to these efforts have also caught the attention of larger organizations. Since messages coming from large companies or political organizations may not garner the trust that a true grass-roots effort might receive, the large organizations have sometimes resorted to creating fake grassroots campaigns. This strategy is often called astroturfing (for its fake grassroots).

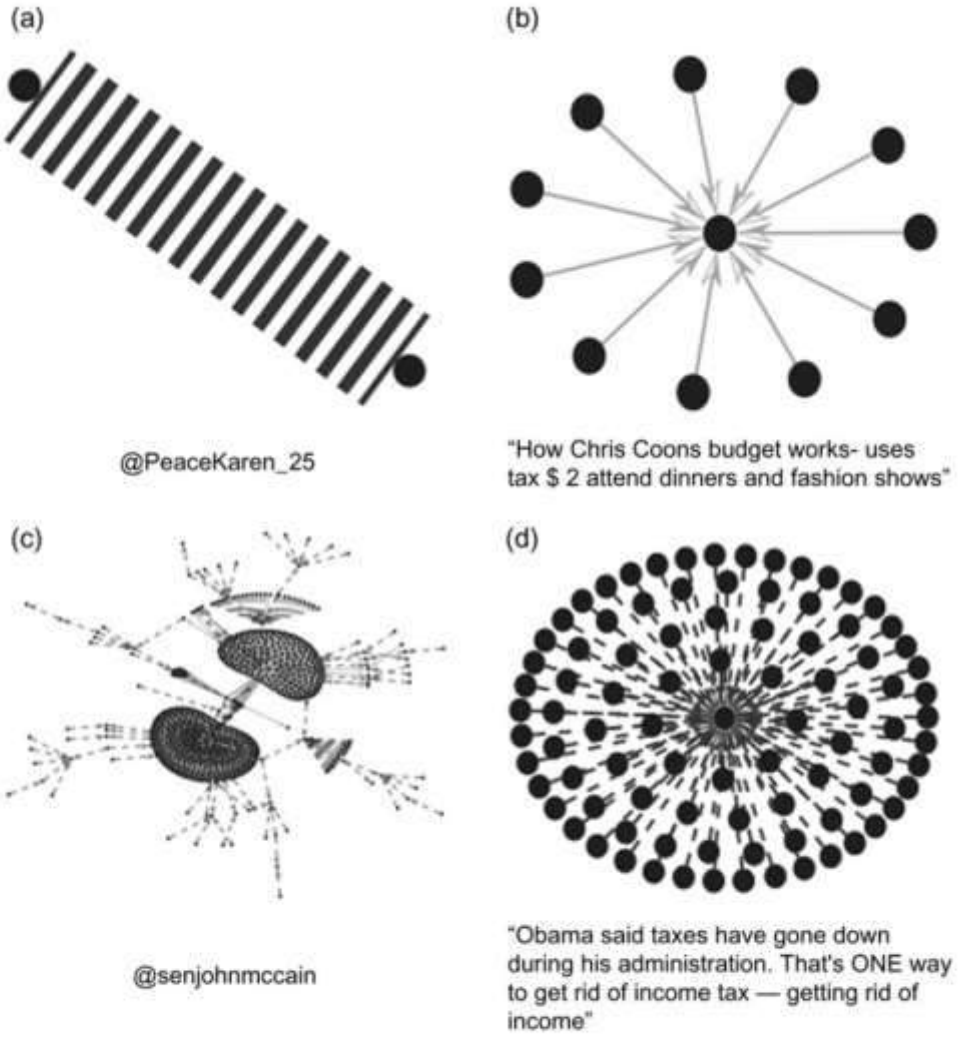

FIGURE 14.3

Examples of Twitter behavior from Ratkiewicz et al. (2011). Dark edges indicate re-tweets and light edges indicate mentions. Graphs (a) and (b) are astroturfing accounts, while graphs (c) and (d) are real accounts.

The technique is not unique to social media. Politicians and public officials have a long history of sending "letters to the editor," written under false names, in which they attacked their opponents or advocated for their own policies. Within social media, astroturfing is relatively common. Researchers have developed a tool called Truthy to detect astroturfing content on Twitter (Ratkiewicz et al., 2011). They present several examples of fake accounts set up on behalf of politicians detected by their system. The analysis includes looking at the social network connections between accounts. For example, Figure 14.3 is taken from their work. It shows two examples of fake accounts and the topics they discussed, together with two examples of real accounts. The difference is in the structure of the social network.

Media investigations have also revealed a deep system of astroturfing. A recent report1 shows that PR firms have "persona management" software that they can use to create an army of fake accounts, prevent them from contaminating one another, and creating suspicious behavior patterns (like those shown in Figure 14.3). The software can automatically create posts and manage accounts, so that a few people

can generate many posts from many accounts. This makes it look like there is a large grassroots movement for a position that is actually advocated for by a large organization.

Analyzing networks to detect legitimate trends versus organized, false accounts is complex. It requires an analysis of social network connections, content analysis, and some detection of the sources behind each account. This type of analysis is not simple, but it provides valuable insights into public opinion and efforts to shape it.

Business Use of Social Media

Measuring Sucess:

A common measure of effectiveness for business is return on investment, or ROI. This number is simply computed as

(Income – Cost)/Cost

Essentially, it measures the fraction or percentage of income earned beyond what was spent. A campaign that costs $1,000 and that earns $1,500 for a company would show a 50% ROI (the company earned back 50% more than it spent).

While this is measurable in some cases, it is not always an easy statistic to compute with social media. Advertising campaigns that combine traditional media with social media may lead to increased sales and show a lot of engagement online, but it is difficult to measure how many of the sales come directly from social media activities. Similarly, if a company begins addressing customer ser vice issues online, that may reduce the number of issues coming in through more traditional channels, like phone calls to a customer service line. However, it is not clear how this impacts income or expenses since people are required to offer service over both channels.

There are other ways to measure the success a company is having in social media. Social media success does not always mean business success, but the following measures can indicate a social media campaign's success.

1. Counts Counting activity is usually quite easy in social media; often the numbers are displayed publicly by the social media site. This may include number of fans, followers, or friends to see the number of people engaged. It may also be number of views on a shared video, number of "Likes" on a post (or similar indications of favorable opinions), or similar counts of people viewing and appreciating content.
2. Social Sharing The counts mentioned in #1 are counting the number of people or their personal actions related to a business's social media site. Sharing is even more important. This could be measured as the number of times an item that the business has posted is shared, the number of times it is mentioned or retweeted on Twitter, or similar counts of sharing behavior.
3. Engagement Rate Counts of people and shares can both be useful, but if a business with one thousand fans gets the same number of shares or likes as a business with one million fans, it indicates that the smaller business is being more successful. Thus, computing the number of engagement activities (likes, shares, etc.) divided by the number of friends, followers, or fans will show how engaged the social media audience is with a business.

4. Interaction For businesses interested in engaging with customers in social media, measuring interactions can be helpful. Counts of the number of customers with whom the business has engaged, the number of conversations, how long each conversation lasts, and how well the interactions are resolved will all indicate how well the business is doing.
5. Referral Rates Often, businesses will use social media to drive people to their websites that are not part of the social media site. Counting clickthroughs, which can be easily measured in server logs or with website analytic software (e.g., Google Analytics), can indicate how much traffic a social media site is driving to the business.
6. Importance and Influence of Users As has been discussed in many places in this text, users vary in their influence and importance in social networks. For businesses, all the measures above treat users identically. In fact, having content shared by more influential users has a much greater impact. Thus, measuring the influence of users can be important. This can be done by computing centrality, if possible, or looking at simpler metrics like number of friends.

Influence is an ideal way to apply many of the social network analysis techniques covered in this book. As an example of these techniques, consider the Twitter network of @frontpageva, a restaurant in Arlington, Virginia. This is a small business with roughly 1,400 followers. Although their account is small compared to those of major corporations, @frontpageva actively uses Twitter to share specials and events and to interact with followers. Social network analysis allows us to see which of their users are most influential and what their reach is.

Beyond the number of followers, we can look at the engagement with @frontpageva on Twitter. Over the course of a summer month when hockey is not in season, they average between 350 and 400 mentions. Dividing that by the number of followers gives an engagement rate of around 30%, which is quite high. They also have roughly the same number of outgoing messages to other people online, indicating that they engage online with others as much as people send messages to them.

Reach is also important, especially for a small business. Figure 15.1 shows the 1.5 egocentric network of @frontpageva on Twitter. Each node is a follower of their account, and size and color indicate the number of followers of each person. Larger, lighter nodes have more followers.

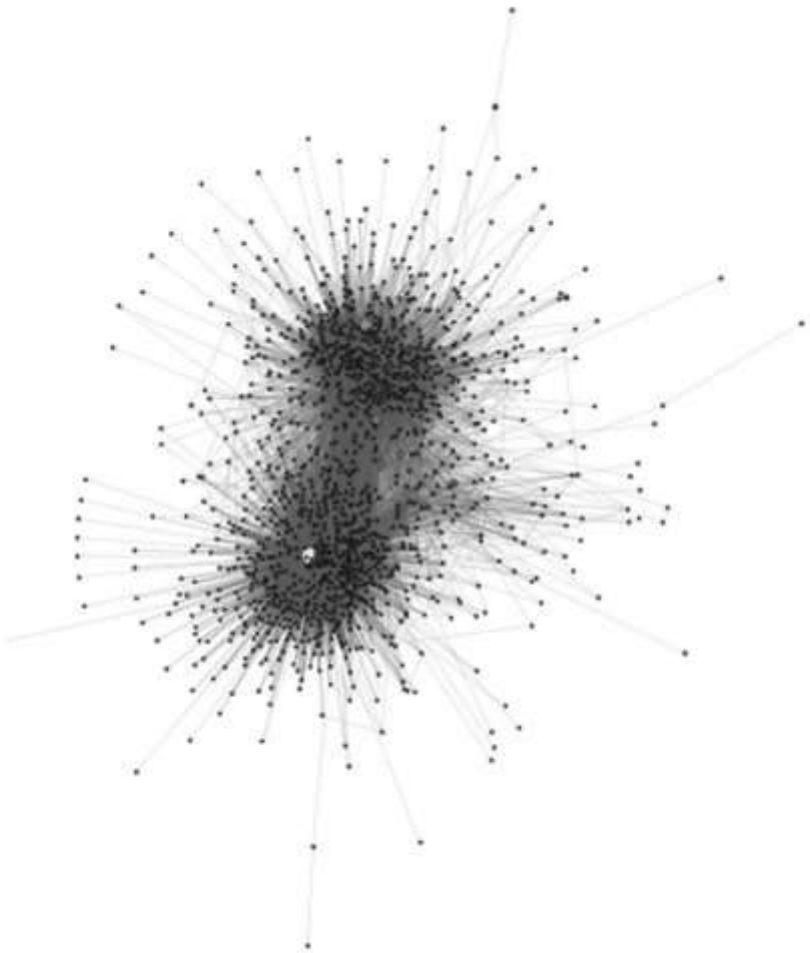

FIGURE 15.1

The 1.5 egocentric network of the @frontpageva account. Larger, lighter nodes have more followers.

Two clusters are apparent in this visualization. Diving deeper to inspect the nodes in each cluster reveals that the top group contains many Twitter accounts that share information about Washington, D.C. and Arlington, Virginia events, parties, and locations. The lower cluster is made up largely of Washington Capitals fans and players because the restaurant is across the street from the Capitals' practice facility.

The large white node in the lower cluster is John Carlson, a defenseman for the Capitals. Because he is a professional hockey player, he has tens of thousands of followers. In the upper cluster, there are a number of large accounts though none as large as John Carlson's node. These represent popular venues and events in the Arlington area.

The @frontpageva account recognizes the importance of these followers. They regularly tweet to John Carlson, and those messages are visible to most of the Capitals fans in the lower cluster since they follow both accounts. John Carlson will also tweet to Front Page, and this online relationship provides valuable exposure and endorsements for the restaurant to their hockey-fan customers.

This example illustrates how many of these metrics can be applied to analyze a business using social media. The next sections will describe examples of excel lent (or poor) use of social media by businesses for specific purposes.

Broadcast example: Will it Blend? Marketing campaign

Blendtec is a manufacturer of high-end blenders. This may not seem like a product that lends itself to highly successful social media campaigns, but the company has one of the most measurably successful efforts of any business.

The company regularly releases humorous videos on YouTube in a series they call “Will It Blend?”1 In the series, the company’s founder demonstrates the blender blending unusual items. These have included butane lighters, a skeleton, Justin Bieber memorabilia, and various electronics. Figure 15.2 shows an example.

FIGURE 15.2

An example of a “Will It Blend?” YouTube video, showing the blender being used on an iPhone.

The videos are not advertised in traditional ways and spread only by viral sharing, yet they have become extremely popular on YouTube. Their top-viewed videos have well over 10 million views each, and their collection of videos all together have 200 million views. The company’s YouTube channel has over 400,000 subscribers.

As mentioned above, counting views or subscribers alone does not necessarily indicate that a social media campaign will help a business. In this case, Blendtec reports a 700% increase in retail sales since launching its YouTube effort.

The videos increased sales by increasing recognition of the brand name and demonstrating the quality of the product. But why do people watch the videos? They are not high-quality productions, they do not feature famous people, and they are not advertised in traditional media.

Blendtec uses several strategies to draw attention. First, the videos show blenders blending things that pique the audience's curiosity; marbles, garden rakes, or guns are not things most people would blend at home. Second, Blendtec tags its videos well and blends items like iPhones that will be interesting to fans of the product being blended. This increases the chance for more views and shares from people who are interested in the items being blended.

In general, Blendtec is successful because people share their videos, and many of them spread virally to millions of people. They get that attention by producing high-quality content that people naturally want to share, and it has had very good results for their business.

Interaction and monitoring example: Zappos customer service

Zappos is an online retailer that sells shoes, clothes, and accessories. It was actively engaged in social media, and in 2009 was named as having the best use of social media by Abrams Research, a company that focuses on social media strategy. Zappos has also been frequently cited for its excellent customer service, and the retailer tries to integrate this reputation into the social media.

Zappos interacts with customers on both Facebook and Twitter. On Twitter, Zappos maintains an account, @zappos service, that answers customer questions and concerns. Representatives for Zappos who monitor the Twitter feed introduce themselves as they change shifts every few hours. A study by STELLAService, a company that studies online customer service, rated Zappos their best performer.3 Over a 45-day period, Zappos was one of only two companies to respond to every service request within 24 hours, and the average response time was under an hour.

Zappos has hundreds of people in its customer service department, but only around 20 handle Twitter requests, and those are in short shifts every day. Usually there are one or two Zappos employees on Twitter at a time, so a very small fraction of the customer service team is needed to manage these requests.

An additional impact of offering customer service online is that other social media users "overhear" these interactions. They can see the back-and-forth conversations, and that helps them get an impression of the company's service.

Zappos serves as an example of how businesses can take advantage of social media to interact with customers. The Zappos service account not only responds to requests sent directly by customers, but it also monitors any posts about Zappos and sends messages to customers with concerns, even if those customers would not have contacted customer support to help with their problem. As a result, upset customers can be reached and helped, even if they would never have asked for help.

Social media failure example: Celeb boutique and the NRA

Measuring success is important, but social media can also have significant impacts on a business's reputation both positive and negative. Social media gaffes are not rare, but businesses want to be extremely careful about them. Particularly because of the interactive foundation of social media, mistakes can be widely shared and backlash can come quickly.

This was evident on July 20, 2012. On that date at a midnight movie showing, an armed gunman opened fire in a crowded theater in Aurora, Colorado, killing 12 people and wounding dozens more. It was one

of the worst mass shootings in history and became the top news item all night and into the following morning.

At 9 A.M., the National Rifle Association(NRA) posted this tweet from one of its most popular accounts:

@NRA Rifleman: "Good morning, shooters. Weekend plans? Happy Friday!"

Later that afternoon, an online clothing retailer, Celeb Boutique, posted the following tweet:

@CelebBoutique: "#Aurora is trending clearly about our Kim K inspired #Aurora dress ;)"

Twitter users were outraged at both posts, inundating the accounts with negative comments. The NRA deleted their tweet three hours later, and eventually apologized and explained that it was posted by someone who had not yet read the morning news. Although this explanation was widely accepted, since the shooting had taken place in the middle of the night on the East Coast, the NRA took another step to mend the situation. Three hours after deleting the post, it deleted the @NRA Rifleman account entirely, losing its 16,000 followers in the action.

Celeb Boutique also removed its tweet after an hour and then issued an explanation that it had not been aware of the shooting when it posted the tweet. This explanation was met with much more skepticism than the NRA's explanation. In addition to the tweet coming much later after the shooting, identifying the reason behind trending topics involves clicking only once on the topic to see the tweets about it. Furthermore, the company is generally a very savvy social media user, interacting personally and well with many people who mention them on Twitter. The "wink" emoticon at the end also led many to believe that they were trying to make an edgy joke. Unlike the NRA, after its initial apology, Celeb Boutique went on to resume its regular tweeting behavior.

What lessons can be drawn from these mistakes? The NRA example basically serves to emphasize that companies should be careful about what they post on social media because inattention to detail can produce a very costly backlash.

The Celeb Boutique case is more complex. The company was taking advantage of a "trending" topic; Twitter identifies the 10 most common words, phrases, or hash tags and marks them trending. Then, anyone can click on those trending terms to see all the tweets about it. Celeb Boutique and many other people and organizations try to include trending content in their tweets to appear when peo ple look at the posts about a popular topic. In this case, "Aurora" was trending because it was the name of the town where the shooting occurred, and Celeb Boutique took advantage of that to market a dress. While connecting a product to popular terms or ideas is often effective in generating traffic as is the case with blending an iPhone in the "Will It Blend?" example connecting a product to a negative idea on social media can generate very negative feelings about a brand.

Conclusions

Social media can be a powerful tool for businesses, but because it reaches so many people, companies must be careful about their posts since mistakes propagate quickly. Social media can be used to broadcast out to users, as in the case of the "Will It Blend?" campaign, to communicate with users, as with Zappos's customer service, or a blend of techniques to receive input and feedback. There are many metrics for measuring success, from simple counts of followers and engagements to measuring ROI.

Many social network analysis techniques covered in this text can be used to identify people of influence within a network, and to guide strategy for reaching out to certain users.

Privacy

There are two major areas to consider in relation to privacy in social media:

how information is shared with other *social media users*, and
how social media websites and services distribute users' information to other *parties*

The main purpose of social media is to share information with other users, but people often want to control who sees what they post. Some information (like reviews, ratings, and comments) is not sensitive or especially personal in nature, and is thus often unrestricted and accessible to anyone. Other information (like photos, personal messages to friends, and contact information) reveals more about people, their relationships, and potentially intimate details. Users often want to restrict who has access to this material, which makes the privacy controls available through the social media website a concern.

A second concern is what a social media website can do with a person's data. It's not uncommon for websites to claim rights to aggregate, share, and sell users' personal data and content. While many issues surround this topic, this chapter will focus specifically on the privacy issues that arise from users' personal data being distributed by a website.

Related to both of these issues is the persistence of information. Once data becomes publicly available on the web, it is archived and cached by many different sites. Thus if users change their minds about what they want to share, previously posted information cannot effectively be removed from the web. This also applies to violations of users' privacy where information is shared without their consent or knowledge.

These trends indicate that controlling the privacy of the personal information we share on social networks is becoming increasingly important. Interest in that information and the associated risks are growing. Understanding privacy first requires understanding the policies of websites, the technology of privacy, and best practices for sharing information online.

Privacy policies and settings

Social media, and social networks in particular, make it much easier for people to share information online. Before these social media technologies became widely available, some people created personal web pages, but the pages were harder to find, and the technical barrier of entry was relatively high for most people. Social media solved these problems by removing these barriers and providing a centralizing location to share and find information.

Privacy settings

People's comfort with sharing online (and the degree to which information is shared with a large audience) has changed over the course of the social media era. Consider as an example the evolution of Facebook's default privacy settings.

In the early days of Facebook, most of a user's shared information was, by default, visible only to the user's friends. Over the course of its life, Facebook's default settings have changed to become increasingly public.

This example explains how visible a user's data is with others online. This visibility is determined by the default privacy settings. As the figures show, the default settings on Facebook have become extremely public over time, with the defaults in mid-2012 allowing everyone on the Internet to see everything a person posts with the exception of their contact information and their birthday.

The number and complexity of privacy management features varies widely between social media sites. Some have no options for privacy. This is especially common on review sites and social bookmarking sites. Large social networking sites (like Facebook and Google +), on the other hand, have many sophisticated tools for controlling privacy, sometimes allowing people to specify lists of individuals who have permission to see each individual piece of information. In between these extremes are sites that offer some limited controls. Twitter, for example, allows users to make their profiles public and visible to everyone, or private and visible only to approved followers.

Table 16.1 shows a matrix of some social media sites and the privacy settings they offer to users.

Table 16.1 Privacy Attributes of Various Social Media Sites

	Social Networking	Mictoblogging	Social Bookmarking	Photo Sharing	Cross-cutting	Location-Based Games	Marketplaces
	Facebook	Twitter	Pinterest	Flickr	Google	FourSquare	Craigslist
How is information collected							
From user	X	X	X	X	X	X	
From other websites (e.g. Facebook, twitter)			X			X	
Information shared by others about you	X	X			X		
Behavioral information (from logs, etc)	X	X	X	X	X	X	
What personal information is collected							
name	X	X	X	X	X	X	
email	X	X	X	X	X	X	X
location	X		X	X	X	X	
photo	X		X	X	X	X	
birthday	X			X	X	X	
posts (updates/text/ photos/etc)	X	X	X	X	X	X	X
How is information used							
registration	X	X	X	X	X	X	X
send email from the registering site	X	X	X	X	X	X	X
customer service	X		X		X		
recommendations (friends, products, etc)	X		X		X		
personalization	X		X		X		
sold							
Who is data shared with							
Other users on website - all			X				X
Other users on the website - user controlled	X	X		X	X	X	
Other internet users (not registered with site)	X	X	X	X			X
Third parties (other companies)	X	X	X	X	X	X	
For analysis provided back to registering site	X	X	X	X	X	X	
For marketing products to you	X			X		X	
For any purpose they choose	X					X	
Law enforcement if requested		X	X	X	X	X	X
aggregated NPII data		X			X		
Companies that have an interest in the registering company	X	X	X	X	X	X	
User control issues							
Accounts/data can be totally deleted	X	X	X	X			
Archived copies kept			X		X		X

Privacy policies

But settings are just one piece of the privacy puzzle; in general, they affect only what other users are able to see on a person's profile. The information collected by sites, how it is used, and how it can be shared with other companies, is rarely controlled through privacy settings. Instead, this is detailed in privacy policies.

Of main concern with regard to privacy policies are the following issues:

What data is collected from users?

In order to establish an account, most websites require an email address and name. Some sites also collect location, photo, birthday, and other data. User's posts are also included here, since it is a type of personal data, but all sites do this since supporting user-generated content is at the core of the sites' functionality.

How is the data collected?

Sites will often collect data from users when they register. Others require users to link to other social networking accounts, like their Facebook and Twitter accounts. This makes it easier to share data on all platforms, and it also provides an additional source from which websites can harvest data about users.

Who is the data shared with?

The data users upload may be shared with other users and other companies. Privacy policies often stipulate if users have control over which specific people can see their data or if it is available to everyone on the site or everyone on the Internet. Policies will also detail which third parties can see the data. These may be companies who do analysis for the hosting website, marketing firms, or other sites that buy the data and use it for whatever purpose they like. The hosting company may give these third parties restricted access, sell their users' personal data, or give some of it away for free.

How is the data used?

Most sites use the personal information users provide to register them for the site and provide communication. They may also use it for customer service, personalizing the users' experiences, making recommendations, or supporting interaction.

What control does the user have?

If a user decides to delete his or her account, what rights do they have to how their data is handled? Some sites will delete all of the user's data and content. Others will keep archived copies for a fixed timeframe or in perpetuity. How account closing and data deletion is handled is usually addressed in the privacy policy.

Privacy policies are generally written in understandable plain English these days an improvement from the times when they were full of legal jargon. Understanding what rights a social media site claims to personal information and content that users create should be an important factor in deciding what information to share.

Some sites have responded positively to actions that their users have under taken in response to their privacy policies. For example, the social bookmarking tool Pinterest originally had a policy that claimed full ownership of any content that users uploaded, including the right to sell any of the images that were

uploaded. This became a major issue for companies and professional photographers who wanted to retain rights to their images. According to these terms, even if someone else uploaded the photographer's image, Pinterest would claim rights to it. After a few months, Pinterest users began strongly objecting to these terms, and Pinterest removed the clause about ownership and the right to sell uploaded images.

Aggregation and data mining

Anonymous use of social media is possible, but remaining anonymous presents serious challenges. Many privacy policies speak of "personally identifiable information." This is data that reveals who you are, like your name or photo. Some data that is not useful on its own (like a ZIP code) may be combined with other data to become personally identifiable. Furthermore, a user may choose to share some information about herself, but not other information. Sophisticated techniques are being developed to allow third parties to infer some of the attributes that users have chosen not to share. This section provides an overview of some research being conducted in this arena.

Deanonymization

A small (but telling) study was conducted by Yates, Shute, and Rotman (2010). The researchers wanted to see how well protected the users' identities really were. This issue is relevant to many bloggers. They selected three anonymous bloggers. These people blogged under pseudonyms and tried to limit the information they shared about their families, places of work, and other personal details. A 2007 study (Qian and Scott, 2007) indicated that over 40% of bloggers censored their posts, including hiding their identities.

For their work, Yates et al. relied on the existence of marketing databases that will sell the name and address of everyone who meets the requestor's demographic requirements. For example, a requestor can specify a set of ZIP codes for location, age ranges, gender, marital status, and type of housing (rental, single family home, etc.). The database is intended for direct marketing and includes over 200 million Americans, with data compiled from a wide range of sources.

Is it possible to discover enough information about anonymous users that their identities could be discovered using a marketing database? The researchers read the blogs looking for the demographic information listed above. Gender was often easy to identify, as was marital status. For some users, a single ZIP code was easy to find because the blogger lived in a less populated area. For others, a set of ZIP codes for the blogger's home city were used. The blog posts also revealed what types of home each person lived in. Because the bloggers often posted about their birthdays, the researchers also found dates of birth for each person.

With this information in hand, the researchers queried the marketing database. Selecting everyone in the ZIP code range who matched the bloggers' age, gender, dwelling type, and marital status, they found that they could uniquely identify each person via their birthday with over 90% accuracy.

The research shows that online anonymity is very hard to maintain because only a few pieces of information which appear to be meaningless for personally identifying someone can be combined to reveal a person's identity.

identifying someone can be combined to reveal a person's identity. If someone is using both anonymous and nonanonymous accounts (e.g., a professional account and an anonymous personal account), more

sophisticate computing techniques exist that can detect this and merge the two identities. These "entity resolution" computer algorithms use a combination of attributes, like addresses or birthdates, structural network data, and other features to merge nodes that represent the same person.

Inferring data

The approaches in the previous section are able to identify people who are anonymous or using multiple accounts. There are other techniques that use data people share in social media to infer more information about them.

One of the first such projects to receive wide media coverage was called gaydar. Developed as a term project at MIT, the application uses Facebook users' friend lists (publicly available by default) to predict the user's sexual orientation. In preliminary experiments, it was able to identify all the known homosexual men in their sample, even though these men had not listed their sexual orientation in their profiles. A similar tool, produced by Stockholm Pride, claims to analyze a person's Twitter posts and provide a "how hetero" score.

Other researchers have used Twitter "following" relationships to identify people's political leanings. Golbeck and Hansen (2010) found the members of the U.S. Congress that a person followed, obtained a score of how liberal or conservative the congressperson was, and combined the scores of the congress people to come up with a score for the Twitter user. Combining users' scores to rate the political preferences of audiences for different media outlets produced results that closely matched previous studies of the media outlet's political leanings. Simple use of public following patterns yielded interesting insights into a user's politics.

Data mining

The studies described so far infer information about specific traits. Data mining, on the other hand, uses many sophisticated computing techniques to discover previously unknown patterns and relationships in large collections of data. Data mining is used in many applications outside of social media as well. For example, one store used data mining on their sales receipts and found that men tended to buy diapers and beer together on Thursdays. Further, they found those families tended to do their main grocery shopping on Saturdays, so the Thursday trips were usually to stock up on things for the weekend. This allowed the store to place a beer display closer to the diapers and ensure that they charged full price on Thursdays.

With social network data, companies will be looking for similar patters. Users with certain attributes may perform certain actions together. This can be used to target advertising to users, or to collect data on those users and sell it to thirdparty companies (so that they can directly market to the users).

Companies are already creating plans to mine social network data and use it in ways that people might not expect. In 2012, Germany's largest credit rating agency which rates how likely people are to repay their loans and thus dictates the interest rates a person might receive on a credit card or their ability to get a mortgage leaked news that it planned to use data from social media to identify potential customers and measure how risky they might be.6 A public outcry about privacy issues shut the project down, but it indicates how information is available through social media.

Recommender systems also use data that people provide to make new suggestions. A person's ratings, reviews, and buying habits are all useful in making suggestions about new items that a user might like.

Some recommender systems also use social data to improve these recommendations. Overall, research shows that users appreciate recommender systems; this example illustrates that technologies that use a person's information need not be threatening or scary.

Data ownership and maintaining privacy online

The interest of companies and organizations in users' data, the trend of social media toward making such information public by default, and the growing number of tools allowing others to discover new information can be overwhelming for social media users. Furthermore, even with well-tuned privacy settings, information shared online can almost never be considered truly private. Many sites have ways for clever or determined people to circumvent the privacy settings. Old data that a user may have deleted may still be archived on other sites. And perhaps the biggest (and technologically simple) threat is the following: Users with permission to see personal information can always copy it and share it with the wrong people.

A user can employ personal strategies to help keep social media data private. However, it is first important to know who owns the data shared through social media. Some websites allow users to own all the data they post. Flickr, the photosharing website, allows posters to retain ownership of everything they share. It also offers options for licenses, so that a user can dictate how others may use their photos. Other websites, like Wikipedia, require authors to give up ownership of their content as soon as it is posted on the site. Facebook technically allows users to maintain ownership of their data, but their terms of service state that you grant them "a non-exclusive, transferable, sub-licensable, royalty-free, worldwide license to use any IP content that you post." That means Facebook is allowed to do anything it wants with the data you upload, including selling it to other people, without paying you anything or asking for your consent.

The Facebook model is common among many social media websites. On one hand, it is important to these companies' business models that they can use people's data. Because most of these sites are free, they need to make money from someone other than the users. Most often, this comes from advertising, particularly from offering advertisers the opportunity to target very specific demographics, based on all the data users upload. In effect, social media users are not the customers of the social media companies; they are the product.

While these business models mean it is unlikely that social media will leave full control of personal data in the hands of users, it does not necessarily mean that the only solution is to stay offline. Understanding the privacy landscape allows users to make better decisions about what (and what not) to share. For example, privacy concerns are rarely voiced around the professional social network LinkedIn. That's largely because the information people put there is not sensitive; it is created for a professional audience, and it is intended to be seen by anyone on the Internet. Users make careful choices about what they post, and they know it will be public.

When using social media for personal rather than professional activities, people can still protect themselves. By default, assume that anything you post could find its way to your boss, potential employers (including jobs you will apply for years from now), all friends, and people who do not like you. Consider the repercussions of the information reaching those people. Then decide which things you are comfortable with reaching a large audience and how much to trust friends to protect those things. Being fully informed about who can see the information, how it can be used, and what the

website's privacy policy is allow users to make the best decisions about what to share. And remember: once content is shared, it can never be fully retracted.