Social Network Data and Predictive Mining (Business Intelligence 2)

8

This chapter covers the second part of our business intelligence discussion and makes the reader learn how organizations can create business value by analyzing social network data. Diverse information about a certain person can be collected from different social media tools and combined into a database to obtain more complete profiles of employees, customers, or prospects (i.e., social engineering). The latter can supplement the social CRM database (see Chap. 5). Particularly, social media may uncover information about what people post, share, or like but also to whom they are connected. By combining or aggregating such information for many individuals in social networks, organizations can start predicting trends, e.g., to improve their targeted marketing (see Chap. 4) or to predict which people are more likely to churn, fraud, resign, etc. Hence, social media are seen as big data in the sense that they can provide massive amounts of real-time data about many Internet users, which can be used to predict someone's future behavior based on past behavior of others. This chapter explains how social networks can be built from social media data and introduces concepts such as peer influence and homophily. The chapter concludes with big data challenges to social network data.

Similar to the previous chapter on business intelligence (Chap. 7), this chapter is mainly situated in the IT department of an organization (i.e., especially regarding the technical execution or implementation of business intelligence techniques). However, to ensure successful business intelligence applications, business input and experience are still required to draw appropriate business decisions. In other words: business intelligence should support the (data-driven) decision-making process (Fig. 8.1).

8.1 Introduction to Social Network Data

In order to introduce the reader to the topic, the use of social network data is first illustrated for targeted marketing before listing applications in other areas.

A. Van Looy, *Social Media Management*, Springer Texts in Business and Economics, DOI 10.1007/978-3-319-21990-5_8

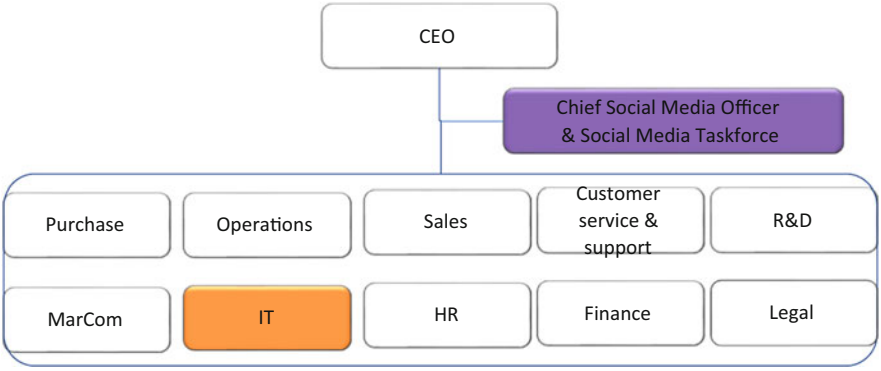

Fig. 8.1 The multidisciplinary approach of social network data

8.1.1 Examples of Social Network Data in Targeted Marketing

In Chap. 4, we discussed the principle of targeted marketing for brands or purchases, which means that Internet users will rather receive those online ads that fit their personal profile (e.g., related to their specific hobbies, interests, profession, the city where they live to receive local deals, etc.). Targeted marketing can be facilitated by investigating personal information revealed by social media tools, e.g., Facebook™ (with information about a user's hobbies, among others), LinkedIn™ (with information about a user's job, among others), or Twitter™ (with information about a user's professional expertise, among others). These different types of personal information can be combined per user (e.g., customer or prospect) into a database for the purpose of deriving different types of clients.

The following examples introduce social network analysis (and predictive analytics in general) for targeted marketing. First, Fig. 8.2 illustrates that if a social media tool reveals that a user (let's say Axl) likes gaming, that user will be more likely to get an online ad for a new game. Similarly for online shops, if browser cookies reveal that a user (let's say Ashley) searched for a specific product in a web shop, that user will be more likely to get an online ad for similar products or similar shops. Hence, targeted marketing can use personal information to better serve people as a (potential) client by trying to predict which products or services the user might be interested in and possibly buy in the near future. However, this kind of (direct) targeted marketing and predictive analytics does not make use of social network data, because predictions of someone's future behavior are limited to past behavior of himself/herself.

On the other hand, targeted marketing can also work indirectly, i.e., via a social network. In this case, it tries to predict which products or services a user might be interested in, namely, (1) based on his/her relationships with other people or (2) based on similar characteristics with other people. For instance, Fig. 8.3 illustrates that if a user (let's say Ashley) bought a new cell phone online, the user's connections in social media tools (e.g., Emma) are more likely to see an online ad for the same cell phone. This indirect way of targeted marketing tries to

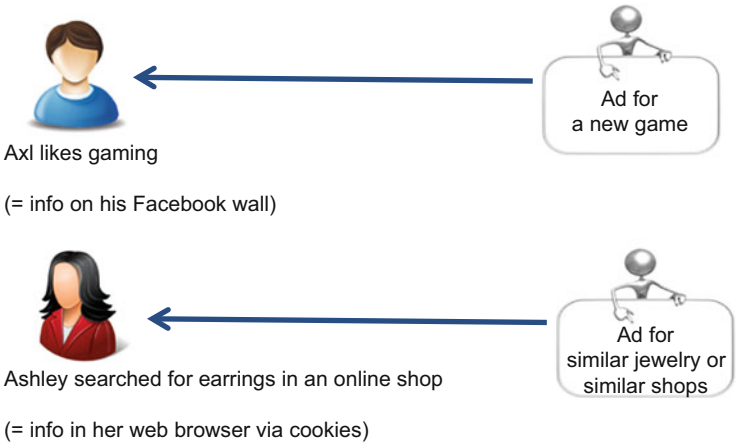

Fig. 8.2 Examples of targeted marketing without social network analysis (direct, online)

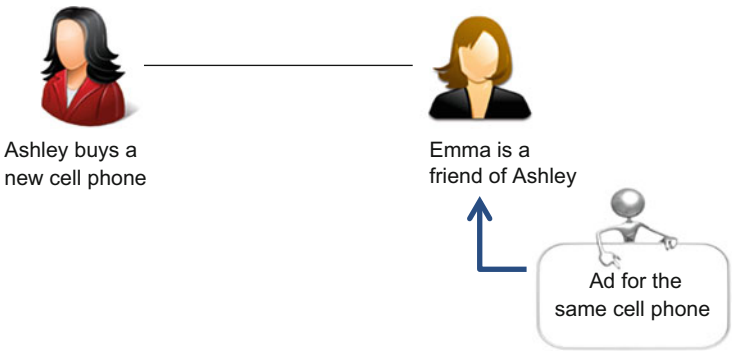

Fig. 8.3 An example of targeted marketing with social network analysis (indirect, online)

estimate the probability that Emma tends to buy the same product as her friend Ashley and which is the topic of the present chapter. In particular, social network analysis tries to predict someone's future behavior based on past behavior of others.

Predictions of future behavior can be based on online information (as seen in the previous examples), but they can also deal with offline information. For instance, shops and supermarkets may give their customers loyalty cards with which they can track the products that are frequently bought (besides personal information, e.g., name and address). Such offline predictions can be made directly, as shown in Fig. 8.4. Another way to predict future behavior is indirectly, i.e., based on a network of client types (Fig. 8.5) which thus requires network data.

A client type is a collection of customers (or prospects) of an organization who conducted similar behavior in the past. It may concern people who regularly buy similar products, such as Axl and Cedric in Fig. 8.5. In this example, Axl and Cedric do not know each other but share similar characteristics by buying the same

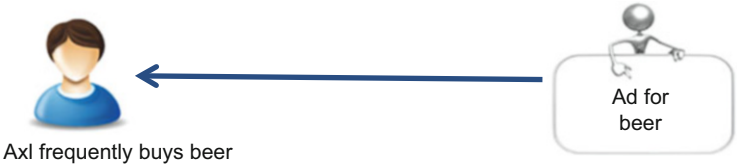

Fig. 8.4 An example of targeted marketing (direct, offline)

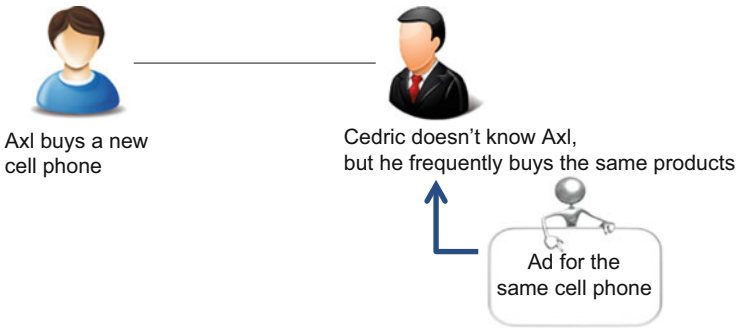

Fig. 8.5 An example of targeted marketing (indirect, offline)

products. Those people are more likely to continue buying similar products in the future.

We note that future behavior can also be predicted based on personal expenses. Ultimately, a bank account may reveal not only personal data (such as a person's name and address) but also which shops a person frequently buys products or services. Banks currently deal with the issue of commercialization of (anonymous) client information, which enables third-party organizations to derive different client types for targeted marketing. For privacy reasons, the information exchange will be anonymous. For instance, a third-party organization will only know people by an anonymous ID number (e.g., persons X, Y, Z instead of Joe, Harry, or Sandy).

8.1.2 Examples of Social Network Data in Other Areas

This chapter explains how social network data can be analyzed in order to discover relevant knowledge for an organization based on profiles with shared characteristics or attributes. Such knowledge discovery can be applied in diverse areas and is not limited to targeted marketing. Some applications for which organizations can use social network data are:

- Customer acquisition and churn prediction (e.g., to predict which customers are more likely to buy or to churn and to send customized recommendations)

- Credit scoring (e.g., to predict which client types are solvable and will likely be able to repay a loan)
- Fraud detection (e.g., to predict which profiles are more likely to fraud)
- Healthcare (e.g., to predict which profiles are more likely to bully or to get a certain disease)
- Other (e.g., stock price prediction, spam detection, counterterrorism, public policy, etc.)

Moreover, network analysis can go further than merely predicting future behavior. For instance, instead of only suggesting interested products to web shop visitors (i.e., targeted marketing), Amazon™ (http://www.amazon.com/) is also investigating how it can proactively send products to a (loyal) customer (i.e., before he/she has placed an order) (Wikipedia 2014). In particular, Amazon™ has a patent on proactive sending, which is called the "method and system for anticipatory packaging shipping" (patent number US 8,615,473 B2 from December 24, 2013). The purpose of proactive sending is to guarantee a fast delivery to increase customer service without necessarily having warehouses in every country. Proactive sending makes use of network analysis and data mining technology to predict which products their customers might buy and when. It can therefore analyze online actions, such as previous orders, the keywords that a customer has used in search engines, websites that he/she has visited, or wish lists in various e-shops that are stored as browser cookies. Additionally, the predictions may rely on what other customers with similar characteristics frequently buy. As such, an organization can collect information of many customers and prospects in big datasets, which can be analyzed with data mining techniques to find patterns or client types and so predict future orders. Amazon™ also intends to send those predicted orders to the warehouse that is closest to a customer's home before he/she orders these products or puts them in a shopping cart.

8.2 Defining Social Network Data

8.2.1 Social Network Modeling Approaches

Research on social networks can be divided into two groups (Provost and Fawcett 2013; Shmueli 2010): (1) descriptive network modeling for social network analysis and community detection and (2) predictive network modeling for link prediction and attribute prediction.

Descriptive network modeling examines social networks to gain insight into the structure of a network and to identify important people or groups. For instance, in the context of viral campaigns (see Chap. 4), centrality measures can be used to detect social leaders, which is useful information for launching a campaign. In particular, descriptive network modeling tries to detect a network (or community) and its members and examines how they are linked to each other, as shown in Fig. 8.6. Figure 8.6 also illustrates that networks can be interconnected to each other

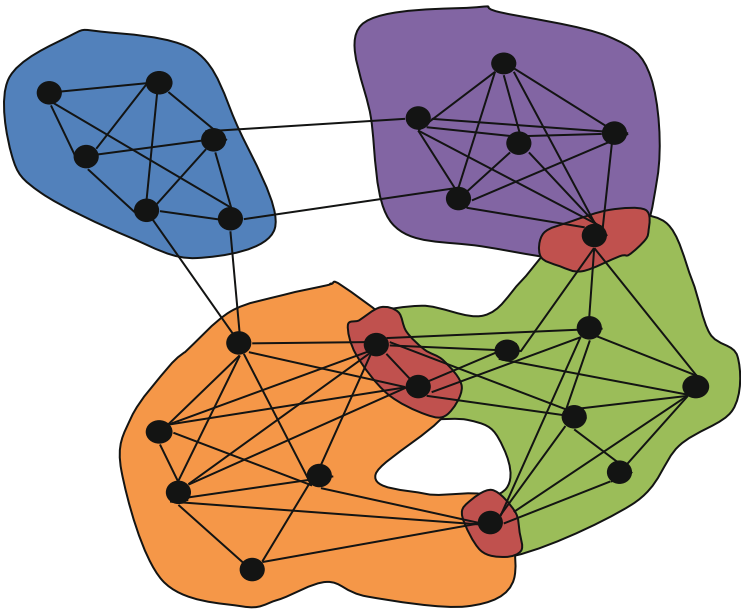

Fig. 8.6 Descriptive network modeling

(e.g., partnerships between organizations). For instance, one network may reinforce the other network by cross selling to each other products and services, similar to the social ripple effect (see Chap. 2). A strong network can also use different channels (e.g., Facebook™, Twitter™, LinkedIn™, blogs, etc.).

On the other hand, predictive network modeling involves link prediction to understand the dynamic changes in a network structure and to predict which connections will be made in the future. Link prediction is frequently used by social media tools to suggest "people you may know" because those people are connected to one or more of your connections (e.g., in Facebook™, LinkedIn™, Twitter™, etc.). Particularly, if more connections are connected to that third person, the more likely it is that you know that person too. Secondly, predictive network modeling involves attribute prediction, which looks for common characteristics or attributes, and is also called "predictive data mining." In particular, it uses historical data (e.g., customer records in a social CRM database; see Chap. 5) to build a predictive model (i.e., a "class probability estimation model") that predicts the unknown value of a class or target variable (e.g., customer acquisition, churn prediction, credit scoring, etc.). Several data mining algorithms exist to automatically build a predictive model, which can be expressed as a mathematical formula (e.g., a linear model for regression or classification) and/or a logical statement (e.g., decision rules). Hence, "mining" literally refers to digging for information to find patterns in big data that can be interpreted in order to distill knowledge that is relevant for a business.

In sum, while descriptive network modeling is often used for a causal understanding of a certain phenomenon (e.g., "How do churning customers typically look like?" or "Why do people churn?"), prescriptive network modeling rather intends to predict or estimate that phenomenon for future use (e.g., "Which other/new customers are more likely to churn in the future?") (Provost and Fawcett 2013).

8.2.2 Definitions

A social network can be defined as a number of persons or a group of persons that are related to each other in an offline or online context (e.g., relatives, friends, colleagues, etc.). Social network analysis or network-based analysis refers to using (i.e., analyzing, interpreting, evaluating) information about links (i.e., connections or relationships) in order to predict future behavior and, for instance, to stimulate selling products, services, etc. Given the popularity of social networks, much predictive power is present in the structure of social networks.

When also social media data are used, the social network is called a "social media network" (e.g., a group of connected users on Facebook™, Twitter™, LinkedIn™, etc.). Social media network analysis refers to a network-based analysis which uses links in social media tools. The rising popularity of social media gives new opportunities to network-based analysis due to the availability of a large amount of new data to be included.

8.2.3 Graph Representation

Social networks are usually visualized in a (mathematical) graph. As shown in Fig. 8.7, it consists of circles or nodes for representing entities (e.g., people, animals, etc.). The lines between the nodes represent the links or relationships between two entities. Consequently, a social network can also be defined as a network of nodes, representing entities, which are connected by links, representing a relationship between two entities (Newman 2010). Subsequently, it will be shown that links may differ in strength, with some links being stronger than others.

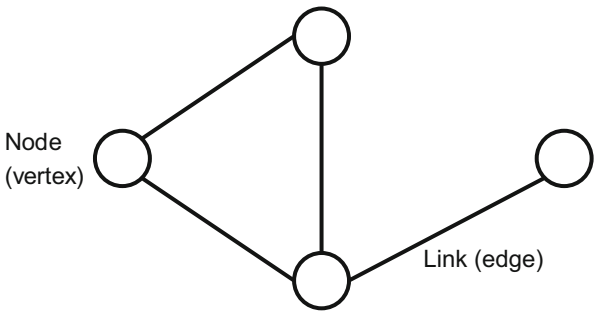

Fig. 8.7 The graph representation of a social network

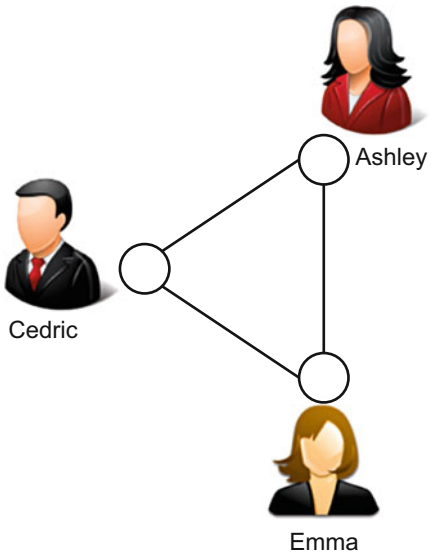

Fig. 8.8 An example of a homogeneous social network

Figure 8.8 illustrates the graph of a social network with people as entities. In this example, the social network consists of three persons, namely, Ashley, Emma, and Cedric. The links between all entities indicate that they might know one another or that they have similar characteristics, i.e., (1) Ashley and Emma are linked, (2) Ashley and Cedric are linked, and (3) Emma and Cedric are linked.

In a social network, it frequently occurs that a relationship between two entities is stronger or weaker than the relationship between two other entities. This strength or weight of a relationship can be indicated by adding a number to the relationship, for instance, 0.2 or 1.1. It is, however, also possible to have different weights between the same entities (e.g., when you feel more connected to your friend than your friend feels connected to you). In such situations, two weights can be assigned by means of unidirectional links and arrows that indicate the direction of the relationship. Figure 8.9 presents the difference between bidirectional and unidirectional links. In sum, bidirectional links can be binary (i.e., they exist or they do not exist) or have a (positive) weight that represents the strength of the relationship, while also unidirectional links exist to form asymmetric relationships.

Figure 8.10 is an example of a graph with bidirectional links. It contrasts to Fig. 8.8 by having one additional entity (called Axl) who is linked to Emma in order to illustrate a heterogeneous social network (i.e., with different types of connections). In this example, the uninterrupted lines indicate coworkers (i.e., between Ashley, Emma, and Cedric), while the dotted line indicates family (i.e., between Emma and Axl). So Emma is simultaneously a relative of Axl and a colleague of Cedric and Ashley. Each link in Fig. 8.10 also has a number that expresses the strength or weight of the relationship. An organization should calculate the weights itself, as the notion of "strength" depends on questions relevant to

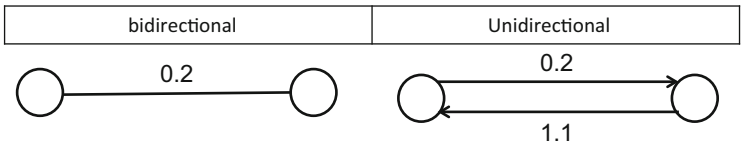

Fig. 8.9 Bidirectional versus unidirectional links in a social network

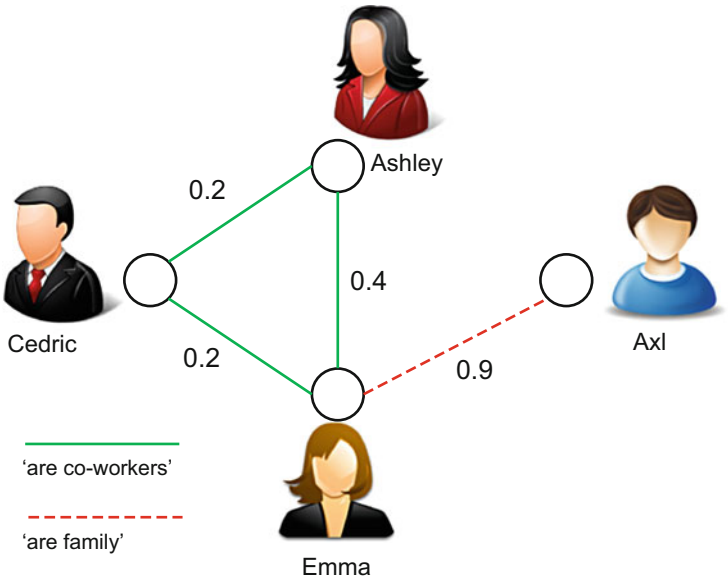

Fig. 8.10 An example of a heterogeneous social network

that organization (e.g., based on the number of products that two entities have in common, the time that two entities stay on the same web page, etc.).

8.3 Social Network Analytics

Before social network data can be analyzed, they must be found and collected in a single database. Boyd and Crawford (2012) give some ideas about how organizations can find datasets to be analyzed as social networks. Besides customer information, the authors mention phone logs, health records, government records, and social media interactions (if available). The unique database under study can thus combine offline big data with online big data.

When using the database for a social network application, a predictive model of the target variable will be found based on other attributes available in the database. Therefore, proper attribute identification and selection is important. For instance, Provost and Fawcett (2013) explain that predictive network modeling is part of

"supervised segmentation" (i.e., by considering a target variable) and tries to iteratively find the most informative attributes on "ever-more-precise" subsets or segments of the database.

The subsequent sections illustrate how a social network analysis can be conducted by looking at concrete applications for reasons of attribute prediction.

8.3.1 Examples of Social Network Applications (with Only Offline Data)

Descriptive and predictive network modeling were originally limited to offline data. For instance, Table 8.1 shows an extract of a customer database.

The final column of Table 8.1 indicates which customers already bought a certain product. In this example, the product was bought by two customers, indicated by the first two rows (i.e., with ID 212 and ID 213). Customer analytics looks for shared attributes or characteristics to predict which other customers (represented by rows) are also likely to buy the same product in the near future. In this example, the customer in the last row of Table 8.1 will probably be interested in the product too as he/she shares many characteristics with the buyers of the product. In particular, customer ID 216 is in the same age category (i.e., age < 30) as the customers who already bought the product and has also placed a high number of orders (order count > 45) of a similar average amount ($50 <$ average amount < 100) and for similar products (i.e., games). These conditions are not met by the other customers in Table 8.1. As customer analytics suggests that customer ID 216 has the same profile as the buyers of the product under study, an organization becomes able to particularly target this customer by proposing a new, personalized offer.

The example shows that general rules can be derived from a dataset in order to predict future behavior of others (e.g., future sales, in this example) and this by identifying people who have a high probability of conducting a certain act (e.g., buying a certain product). Deriving such general rules from a dataset in order to create knowledge is called "mining" or "profiling."

Table 8.1 An example of customer analytics based on historical data

ID	Age	Last order	Order count	Average amount	Order interests	Promo	Product purchase
212	**25**	14/05/31	**50**	**85**	**Games**	**X**	**Yes**
213	**22**	14/02/28	**65**	**73**	**Games**	**X**	**Yes**
214	45	14/09/15	12	123	Books, beauty	Y	No
215	50	14/08/17	5	230	Books, beauty	None	No
216	**18**	14/09/04	**66**	**55**	**Games**	**X**	**No**

The telecommunications sector frequently applies social network analysis based on offline data, e.g., who is calling who and for how long (Pinheiro 2011; Verbeke et al. 2014). For instance, for reasons of customer acquisition, a telecommunications organization may wish to answer the following question: "Given that a customer named Axl has bought a certain telecommunications service, what is the probability that Axl's friends will buy the same service too?" The social network for this example can be a network based on the phone calls made, in which (1) the nodes are represented by customers or prospects, (2) the links are based on phone calls with a minimum duration (let's say of at least 10 s), and (3) the weights are based on the aggregate of all phone calls made between the nodes. The target variable is to know who is more likely to be interested in the service and who might be less interested. Variables to predict this target variable may relate to geographical and demographic data, the level of technological expertise, financial information, and (most importantly) the first-degree connections. In other words, Axl's connections are more likely to buy the same service, and particularly those with similar characteristics on the other variables or attributes (e.g., geography, demography, technological expertise, finances). The resulting predictions will facilitate targeted marketing. Afterwards, the social network can be evaluated by verifying how many of the predictions actually turned into sales.

A similar social network can be used for customer churn prediction, which is a relevant business issue because keeping existing customers satisfied is much more difficult than acquiring new customers (see social CRM, Chap. 5).

8.3.2 Examples of Social Media Network Applications (with Online Data)

Social media data can be added to social networks in order to enhance the dataset. Particularly, social media may supplement corporate data to help find the links (i.e., relationships or connections) between persons. This input is particularly useful for those organizations that do not have their own network data, such as a network of phone calls in the telecommunications sector.

Consider the following example, in which social media reveal connections between users:

-  is dad of  .
-  is a friend of  and  .
-  and  are colleagues.

This social media information can result in a social network, as presented in Fig. 8.11.

Furthermore, social media tools (e.g., Facebook™) do not only record the people who are known by a user (as connections) but also the user's posts and the pages and posts that the user "likes" or "shares." Such personal data can be added to the user's profile, which stimulates social media tools to commercialize their data (i.e., to sell social media data to third-party organizations).

As social media may quickly result in an explosion of data, organizations usually consider the links of the direct connections only (i.e., the first-degree connections, instead of connections of connections of connections, etc.).

Figure 8.12 illustrates how online data can create a quasi-social network, starting from Internet users (browsers) visiting web pages with UGC (user-generated content; see Chap. 2).

For instance, since Alex and Ann visited the Facebook™ page of John, it can be derived that they are both linked to John. Also Alex and Ann will probably be connected, as Alex and John visited the Facebook™ page of Ann.

Furthermore, it can be derived that Ann, Pete, and Jeff are connected, as they visited the same web pages (i.e., OnlineReviews.com, OrganizationBlog.com, and the Facebook™ page of Company XYZ).

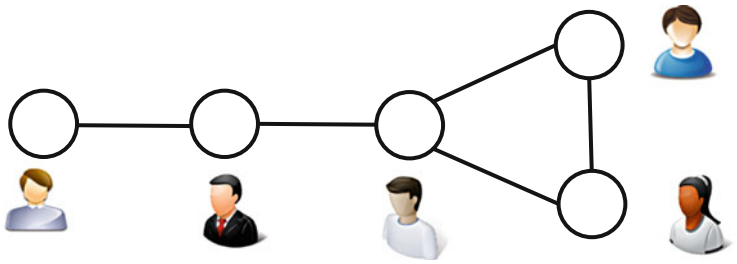

Fig. 8.11 An example of a social media network

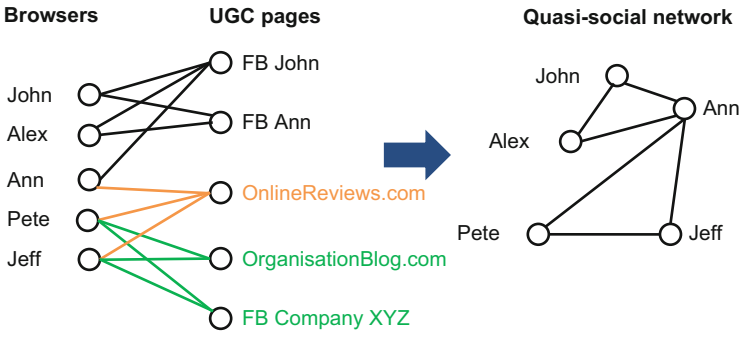

Fig. 8.12 An example of a quasi-social network

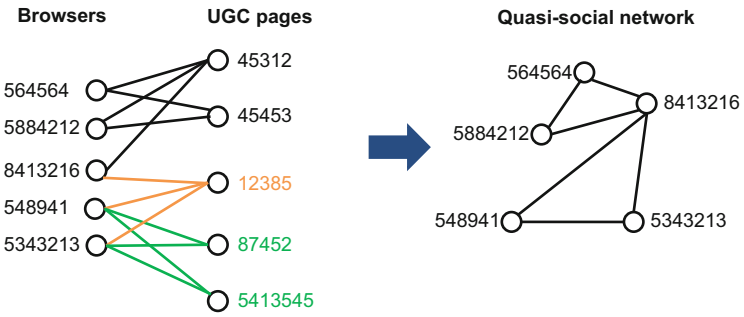

Fig. 8.13 An example of a quasi-social network, made anonymous

The result is a "quasi-social network," because the direct connections are based on assumptions derived from variables (i.e., co-visitations of the same web pages). Nonetheless, it remains unsure whether the persons involved actually know each other in real life, which explains the prefix "quasi-."

The specific procedure to create a quasi-social network is called "bipartite graph mining," because it aims to find bipartite relationships, i.e., a graph with two types of nodes (e.g., browsers versus UGC pages or people versus products, interests, Facebook™ likes, etc.).

Predictive mining tries to protect the privacy of social media users by de-identifying both the browser names and the UGC pages by means of (random and unique) ID numbers, as shown in Fig. 8.13. Hence, the network is "double de-identified," namely, (1) the ID numbers remain anonymous, and (2) no information on browsers or pages is saved.

The quasi-social network can be improved by adding weights to the links. In this example, a weight or strength refers to the number of UGC pages that a node has in common with its direct neighbors. The more pages that are co-visited by two browsers (or individuals), the higher the weight will be. For instance, the link between browser 564564 and browser 5884212 has a strength of 2 (i.e., they co-visited two UGC pages, namely, page 45312 and page 45453), whereas the link between browser 564564 and browser 8413216 has a strength of 1 (i.e., they co-visited one UGC page, namely, page 45312). The former link is thus stronger than the latter, based on the number of co-visitations.

The next section clarifies how a quasi-social network such as Fig. 8.13 can be used for diverse applications.

8.3.3 Mining Algorithm

In order to make predictions from a (quasi-)social network, a mining algorithm automatically runs to process the massive amount of big (social) data. This section illustrates step by step how an algorithm typically works by means of a case study

(Minnaert 2012). Hence, although an algorithm normally runs automatically, the different steps are subsequently discussed to explain the process. The case study is as follows.

Assume a situation in which an online ad tries to convince people to act (let's say to buy a product).

You have data from different groups of people that constitute a network, namely:

- People who took action after seeing the ad (i.e., the buyers)
- People who saw the ad without taking action (i.e., the nonbuyers)
- People who have not yet seen the ad

Try to predict which persons in the third group are more likely to take action after seeing the ad (i.e., to become buyers) in order to improve targeted advertising.

PS We remind the reader that a similar case study can be conducted for other applications, such as credit scoring, fraud detection, spam detection, predictions regarding stock prices or health issues, etc.

Figure 8.14 shows the start situation of the case study. Suppose that there are:

- Two buyers (i.e., represented by $a1$ and $a5$)
- Two nonbuyers (i.e., represented by $a3$ and $a4$)
- Four unknown nodes (i.e., represented by $a2$, $a6$, $a7$, and $a8$)

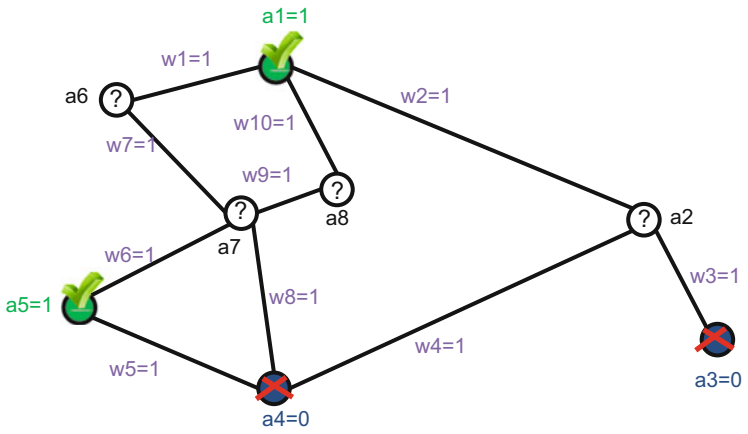

Fig. 8.14 Case study of a mining algorithm (start situation)

Each of these nodes has a unique ID number, namely determined by clockwise numbering in Fig. 8.14. Each known node also indicates a certain action, namely:

- 1 for action (i.e., buying the product after viewing the ad, e.g., $a1 = 1$)
- 0 for nonaction (i.e., not buying the product after viewing the ad, e.g., $a3 = 0$)

Furthermore, the links between the nodes have a strength or a weight (w). In order to facilitate the example, all connections have an equal strength or a weight of 1 ($w = 1$).

The final aim is to infer a probability distribution for the unknown nodes (i.e., indicated with a question mark) in order to identify which of the unknown nodes are more likely to act. Or in other words, which ones are potential buyers and which ones are rather not? For this purpose, we need to derive a score for the actions of the unknown nodes, as was already done for the buyers with a score of 1 and for the nonbuyers with a score of 0. Consequently, the probability to be derived will range between 0 and 1, with scores closer to 1 referring to a higher probability of buying a certain product after seeing a particular ad. In fact, probability values always range between 0 and 1.

Figure 8.15 shows the basic prediction, which is the average probability derived from all known "seed" nodes:

- (# buyers)/# seeds $= 2/4 = 1/2 = 0.5$
- $0 <$ probability < 1

Seeds are the technical term for referring to the known nodes. In this example, there are two buyers and two nonbuyers, which makes four seeds. As half of the seeds bought the product, the basic prediction results in temporary probabilities of 0.5 (i.e., the number of buyers—which is 2—divided by the number of seeds, which

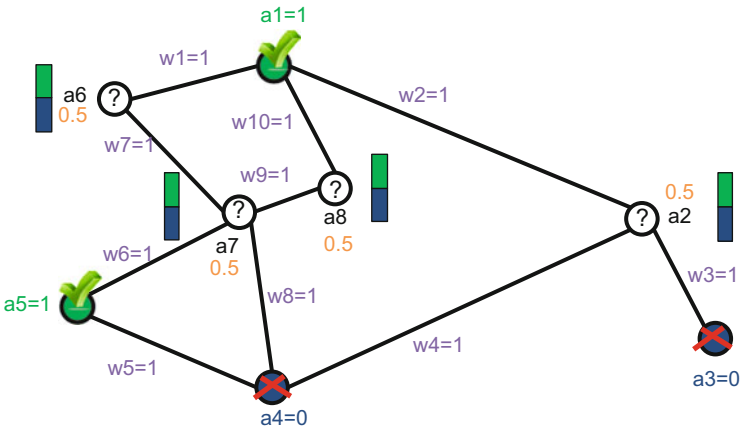

Fig. 8.15 Case study of a mining algorithm (basic prediction)

is 4). These temporary probabilities will be refined throughout different iterations and will be reused as direct input for the next iteration by means of a rectangle (as shown in Fig. 8.15). The rectangles have two parts, with the upper part indicating the likelihood of buying and the lower part indicating the likelihood of nonbuying for a specific node. Hence, after the basic prediction, the likelihood for all unknown nodes is still 50–50, and the upper part and lower part of the rectangles have equal sizes.

In a first iteration, the basic probabilities are refined by calculating the average probability over all direct neighbors of a particular unknown node.

$$[\text{sum of}(\text{neighbors*their weights})]/(\text{sum of their weights})$$

As all weights in this example are equal to 1, the calculation is just the sum of the direct neighbors, divided by the sum of their weights. The new probabilities are shown in Fig. 8.16. For instance, node $a2$ has three direct neighbors, namely, one buyer (i.e., $a1$) and two nonbuyers (i.e., $a3$ and $a4$). Hence, the sum of the direct neighbors is 1, divided by 3 (or 1 out of 3 was a buyer). This calculation results in a probability of 0.33 for node $a2$. The same reasoning applies to calculate the new probabilities for the other unknown nodes, with $a6$ having 2 direct neighbors, $a7$ with 4 direct neighbors, and $a8$ with two direct neighbors. Remember from Fig. 8.15 that the basic probabilities for all unknown nodes is 0.5, which should be taken into account when calculating the new probabilities for $a6$, $a7$, and $a8$.

In a second iteration, the rectangles and probabilities are adapted to the results of the previous iteration. Figure 8.17 shows that the starting value for $a7$ is again 0.5, whereas the starting value for the other unknown nodes has changed (i.e., to 0.33 for $a2$ and 0.75 for $a6$ and $a8$). The same calculation can be redone as in the previous iteration (i.e., the sum of the direct neighbors divided by the sum of the weights), but taking into account the predictions from iteration 1. Figure 8.17 specifies that the new probabilities in iteration 2 do not change for $a2$, $a6$, and $a8$, because the direct neighbors did not change in value throughout iteration 1. Only a new prediction is needed for $a7$, resulting in a probability of 0.625.

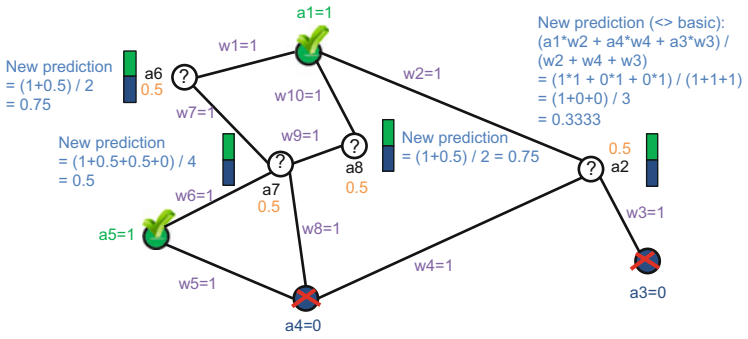

Fig. 8.16 Case study of a mining algorithm (iteration 1)

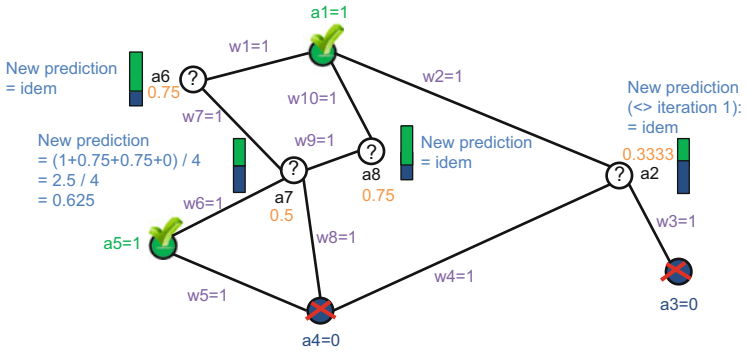

Fig. 8.17 Case study of a mining algorithm (iteration 2)

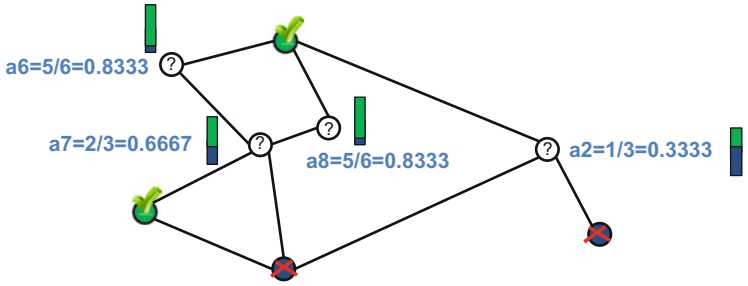

Fig. 8.18 Case study of a mining algorithm (iteration x—end situation)

We continue iterating until the probabilities remain stable. After a certain number of iterations (let's say x iterations), the probabilities of the unknown nodes will not change anymore and the case has reached an end state (Fig. 8.18).

In real life, the calculations happen automatically by means of an algorithm, but this manual example illustrates how organizations can handle network data to gain knowledge. The final probabilities for the unknown nodes represent a value or a score between 0 and 1 (i.e., similar to a percentage). Be aware that a probability of 0.33 should not be read as "a chance of 1 out of 3 to buy a product" but rather as a percentage. The goal is to compare the probabilities in order to decide which persons in the third group are potential buyers and are worth targeting in the organization's marketing campaign.

For instance, advertisers can take 10 % as a bottom line (i.e., as an under limit), which represents the percentage of nonaction. In other words, they will then target those unknown nodes with a probability between 0.9 and 1. As Fig. 8.18 does not show an unknown node with a value of 0.9 or higher, none of the unknown nodes in our example will be targeted if the bottom line is set at 10 %. On the other hand, when the bottom line is 50 % or 0.5, for example, all unknown nodes except for $a2$ (which has a probability of 0.33) will be targeted for the ad.

One of the reasons why the previous example is relevant to organizations is because online ads are expensive (see Chap. 4). Therefore, organizations might profit from only showing their ad to those people who seem more responsive to the message of the ad. As explained in Chap. 4, when people (Internet users) navigate to a website, their browser cookies become available to that website. If that website also sells space for ads to a central ad network (e.g., DoubleClick Ad Exchange), a real-time bidding process starts in which organizations with ads that correspond to a user's interest will bid higher. With the business intelligence techniques for predictive mining, an organization can predict which ad must be shown to which user.

For the purpose of predictive mining, an organization can create and combine different databases, for instance, (1) a database of Internet users who clicked on the organization's ad to collect information about his/her browser, IP address, and cookies (e.g., to uncover his/her interests based on previous clicks, visits of other websites, etc.) and (2) a customer database with information about the buyers of the product or service in the ad (e.g., whether or not they have seen the ad before the purchase). Two browsers can, for instance, be assumed to be connected as neighbors (or quasi-friends) if they have visited the same websites or websites with similar content.

The study of Provost et al. (2009) about online brand advertising shows that organizations best target those people who are part of the 10 % best ranked nodes (thus with a probability of 0.90) in order to have good profit from an ad. In particular, the authors noticed a lift in brand actor density (i.e., real buyers) for a bottom line of 10 %.

8.4 Triggers for Social Network Data

The previous section clarified that direct neighbors in a social network are targeted in order to make predictions. This method works well because social network mining exploits peer influence and homophily, which concerns two important triggers for mining and business intelligence in general. Peer influence applies when you know the other persons in your network, whereas homophily considers strangers with similar characteristics.

8.4.1 Trigger 1: Peer Influence

The first trigger for targeting network neighbors is peer influence (Aral and Walker 2011). This trigger can be illustrated by wondering about the question: "If my friends jump off a cliff, would I jump too?" The question reflects on the degree to which a person will be affected by its peers in a specific situation. Thus, peer influence refers to how the behavior of one's peers can change the likelihood that a person will engage in a certain behavior. For instance, people can consult relatives and friends before purchasing products (e.g., home electronics).

For an organization, it is important to know the influencers in a social network because behavior can cascade from one node to another (similar to an epidemic or contagion). Peer influence is present in situations related to opinions or rumors, but it can also explain situations related to public health, failures in financial markets, etc. The success of viral campaigns, such as the Hotmail™ campaign in the 1990s (see Chap. 4) can be explained by peer influence as emails are usually sent to people you know. Hotmail™ added a sentence at the bottom of each outgoing mail ("PS: Get your private, free email from Hotmail™ at http://www.hotmail.com"), which resulted in about 12 million new Hotmail™ users within 18 months. Another example relates to book publishers who may rely on peer influence to reach higher sales by giving free copies to influential readers when a new book is released. In sum, peer influence occurs when people in the same network have attributes (e.g., gadgets), and they are no total strangers to each other.

Let's look again at Fig. 8.10, which illustrates a social network with coworkers and family. Suppose that three network neighbors (namely, Ashley, Emma, and Cedric) have a cell phone of the same color and of the same brand, while one other connection (Axl) has a cell phone of another color and of another brand. Axl is connected to only one node in the network, which explains why peer influence is less powerful for him.

8.4.2 Trigger 2: Homophily

Another trigger for targeting network neighbors is homophily, which explains how social relationships arise (McPherson et al. 2001). Homophily means "love of the same" or the tendency of individuals to associate and bond with similar others. This trigger can be illustrated by the expression: "birds of a feather flock together" which indicates that similarity breeds connection. The expression literally refers to groups of birds that pass by. One group always covers the same type of birds (e.g., geese with geese, ducks with ducks, or sparrows with sparrows). It is hard (or even impossible) to find a mixed group of, let's say, sparrows and geese.

Translated to human beings, people with the same characteristics are more likely to form a network (e.g., based on similarities in age, gender, class, values, beliefs, etc.). The phenomenon of homophily regularly takes place because similarity makes communication and relationship formation between people easier. For instance, homophily often leads to homogamy, which is a marriage between people with similar characteristics. Hence, homophily explains why people tend to behave similarly or buy similar products and services, even if they are total strangers to each other. The more characteristics they share, the more likely it is that they will conduct similar behavior.

Figure 8.19 illustrates an example of homophily (Minnaert 2012). Assume that the three persons described in Fig. 8.19 sit in the same bar and do not know each other. Who will be more likely to start talking to each other?

All three persons described in Fig. 8.19 are men in the same age category. The two men on the right seem to have a link based on their job and children, whereas

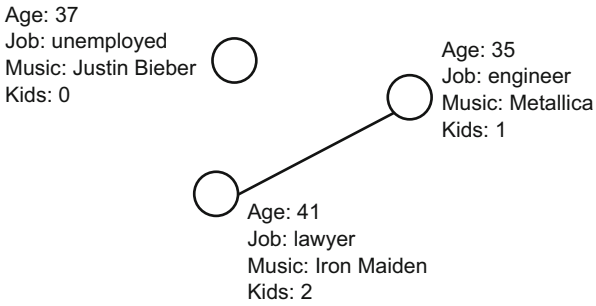

Fig. 8.19 An example of homophily

the person on the left is unemployed and without kids. Consequently, based on homophily, we can assume that the two men on the right are more likely to influence each other and spontaneously start talking to each other in the bar.

8.4.3 Peer Influence Versus Homophily

Peer influence and homophily can coexist, but one trigger will generally become more decisive to act (or not to act) than another in a specific situation. The effect of homophily also remains longer than peer influence, as the former is based on intrinsic similarities.

As an illustration, think about a situation in which obese children are playing together. Can this situation be described by mainly peer influence or mainly homophily? If the former is true, then obesity is considered as being contagious, meaning that a child will gain weight when playing with obese friends. Hence, the situation of obese children playing together is better explained by homophily, meaning that people with similar characteristics are more likely to become friends and to play with each other. Some degree of peer influence plays a role as friends know each other instead of being strangers. For instance, based on peer influence, one friend can convince another friend to eat more cookies or to drink more soda during a school break. Nonetheless, the degree of homophily is stronger in this situation and is thus the main trigger (e.g., because obese children can be bullied by other children).

Although peer influence and homophily are complementary (and can be present simultaneously), viral marketing generally profits more from high peer influence (e.g., for sharing videos), while direct marketing (e.g., online ads) benefits more from high homophily.

Further on, both triggers can be translated to the theory on the diffusion of innovations (Rogers 2003), which we discussed in Chap. 1. The so-called innovators and early adopters are more likely to have other adopters in their neighborhood because of similar characteristics (i.e., homophily) and will be less influenced by peers. However, also a temporal effect of peer influence can be

created, particularly for the "late majority" and "laggards" who are more likely to adopt (or not to adopt) based on peer influence rather than homophily.

What about quasi-social networks? The study of Provost et al. (2009) explains that a particular connection becomes stronger when Internet users (browsers) navigate more to the same web pages (frequency), unless it concerns popular pages to which many people navigate (popularity) and unless it concerns older visits (in the past). This implies that mainly homophily counts for creating quasi-social networks between users. Peer influence is only present if the users also know each other. Although social media tools can give insight into which users seem to know each other (e.g., as being connections on Facebook™), the degree to which connections know each other remains unsure.

8.5 Big Data Challenges to Social Network Data

This chapter has shown how social network data (as big data) can create business value for organizations by predicting future behavior. The research field of big social data, however, still faces some important challenges (Boyd and Crawford 2012; Manovich 2012).

First of all, social media give the opportunity to add a lot of data to a network analysis, but it rather concerns personal data (e.g., about interests, hobbies, professions, connections, etc.). In the end, privacy remains a big challenge, and efforts have been made to anonymously analyze big data.

Other big data challenges relate to the identification and analysis of social network data. Regarding the former, collecting and getting access to social media data are frequently an issue for organizations. This explains why some sectors make more use of social network data than others. For instance, predictive mining is frequently done in the telecommunications sector which has access to organization-owned datasets with customer information and lists of phone calls, text messages, mobile data, etc. (Pinheiro 2011; Verbeke et al. 2014). Regarding the analysis of social network data, tools and algorithms must be available that are able to accurately deal with the massive amount of big social data. Technical issues still exist to maximize the computational power (e.g., memory issues of computers) and algorithmic accuracy (e.g., to increase the accuracy of predictions made by algorithms). Furthermore, research may continue to enhance knowledge about visualizing social networks, taking into account the impact of peer influence and homophily.

Consequently, predictive mining can give new opportunities to organizations, but this research field is still open for improvement to facilitate its practical use.

8.6 Takeaways

Predictive mining refers to finding similar characteristics of people in a dataset (e.g., in a customer database and/or in social media data) in order to predict future trends or future behavior of others. Social CRM can be used as input for a customer dataset under study (see Chap. 5), among others.

As the predictions are based on profiles (or types of anonymous people with similar characteristics) instead of individuals, predictive mining can also be called "profiling." These business intelligence techniques are frequently used within the context of online advertisements (see Chap. 4), as it facilitates targeted marketing. For instance, if you know that a certain person already bought a specific product or service, what is the probability or likelihood that his/her connections or other people with similar characteristics will be interested in buying the same product or service? But applications of predictive mining are not limited to targeted marketing and can be found in diverse areas, such as predicting customer acquisition and churn, credit scoring, fraud detection, counterterrorism, public policy, healthcare, etc.

Social media have drastically changed the way big data are used. They do not only provide organizations with new data (e.g., about interests, hobbies, professions, connections, etc.) but also allow a more personalized approach due the additional insights that can be gained from social media data. While the previous chapter focused on big (social) data analytics for text mining, this chapter explained why big data analytics is important to predict relevant trends based on peer influence and homophily. Although social media data are rather new and network analysis requires an effort, organizations can already make use of social network data with promising results. In conclusion, a further examination of social network data is worthwhile to overcome current challenges in this research field and to give new opportunities to contemporary organizations in order to perform better.

8.7 Self-Test

- What does social network mining mean?
- Can you explain how social networks are mathematically represented?
- Can you interpret the results of social network mining? For instance, can you give advice to whom an organization best shows a targeted ad in order to save money? Please motivate your choice.
- Are you able to distinguish the degree of peer influence from homophily for a given case description? Think about different situations in which you buy products or services, and evaluate to which degree such situations can be described by homophily and/or peer influence.
  - In a supermarket, travel agency, clothing shop, bank, insurance company, etc.

Bibliography

Aral, S., & Walker, D. (2011). Identifying social influence in networks using randomized experiments. *IEEE Intelligent Systems, 26*(5), 91–96.

Boyd, D., & Crawford, K. (2012). Critical questions for big data. *Information, Communication and Society, 15*(5), 662–679.

Manovich, L. (2012). Trending: The promises and the challenges of big social data. In M. K. Gold (Ed.), *Debates in the digital humanities* (pp. 460–475). Minneapolis: University of Minnesota Press.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology, 27*, 415–444.

Minnaert, B. (2012). [Guest lecture of Bart Minnaert in the course Creating Value Using Social media at Ghent University, November 2012].

Newman, M. (2010). *Networks: An introduction*. New York: Oxford University Press.

Pinheiro, C. A. R. (2011). *Social network analysis in telecommunications*. Hoboken, New Jersey: SAS Institute and Wiley.

Provost, F., Dalessandro, B., Hook, R., Zhang, X., & Murray, A. (2009). Audience selection for on-line brand advertising: privacy-friendly social network targeting. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*

Provost, F., & Fawcett, T. (2013). *Data science for business. What you need to know about data mining and data-analytic thinking*. California: O'Reilly Media.

Rogers, E. M. (2003). *The diffusion of innovations* (5th ed.). New York: Free Press.

Shmueli, G. (2010). To explain or to predict? *Statistical Science, 25*(3), 289–310.

Verbeke, W., Martens, D., & Baesens, B. (2014). Social network analysis for customer churn prediction. *Applied Soft Computing, 14*, 431–446.

Wikipedia. (2014). *Amazon.com. Fulfilment and warehousing*. Retrieved July 16, 2014, from: http://en.wikipedia.org/wiki/Amazon.com#Fulfillment_and_warehousing