

This chapter is from *Social Media Mining: An Introduction*.  
By Reza Zafarani, Mohammad Ali Abbasi, and Huan Liu.  
Cambridge University Press, 2014. Draft version: April 20, 2014.  
Complete Draft and Slides Available at: <http://dmml.asu.edu/smm>

# Chapter 10

## Behavior Analytics

What motivates individuals to join an online group? When individuals abandon social media sites, where do they migrate to? Can we predict box office revenues for movies from tweets posted by individuals? These questions are a few of many whose answers require us to analyze or predict behaviors on social media.

Individuals exhibit different behaviors in social media: as individuals or as part of a broader *collective behavior*. When discussing *individual behavior*, our focus is on one individual. *Collective behavior* emerges when a population of individuals behave in a similar way with or without coordination or planning.

In this chapter we provide examples of individual and collective behaviors and elaborate techniques used to *analyze, model, and predict* these behaviors.

### 10.1 Individual Behavior

We read online news; comment on posts, blogs, and videos; write reviews for products; post; like; share; tweet; rate; recommend; listen to music; and watch videos, among many other daily behaviors that we exhibit on social media. What are the types of individual behavior that leave a trace on social media?

We can generally categorize individual online behavior into three categories (shown in Figure 10.1):

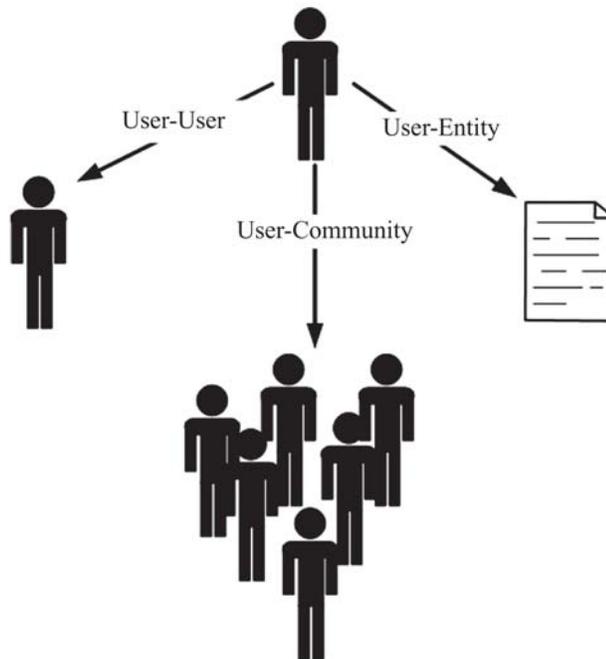


Figure 10.1: Individual Behavior.

1. **User-User Behavior.** This is the behavior individuals exhibit with respect to other individuals. For instance, when befriending someone, sending a message to another individual, playing games, following, inviting, blocking, subscribing, or chatting, we are demonstrating a user-user behavior.
2. **User-Community Behavior.** The target of this type of behavior is a community. For example, joining or leaving a community, becoming a fan of a community, or participating in its discussions are forms of user-community behavior.
3. **User-Entity Behavior.** The target of this behavior is an entity in social media. For instance, it includes writing a blogpost or review or uploading a photo to a social media site.

As we know, *link data* and *content data* are frequently available on social media. Link data represents the interactions users have with other users, and content data is generated by users when using social media. One can

think of *user-user* behavior as users linking to other users and *user-entity* behavior as users generating and consuming content. Users interacting with communities is a blend of linking and content-generation behavior, in which one can simply join a community (linking), read or write content for a community (content consumption and generation), or can do a mix of both activities. *Link analysis* and *link prediction* are commonly used to analyze links, and *text analysis* is designed to analyze content. We use these techniques to analyze, model, and predict individual behavior.

### 10.1.1 Individual Behavior Analysis

Individual behavior analysis aims to understand how different factors affect individual behaviors observed online. It aims to correlate those behaviors (or their intensity) with other measurable characteristics of users, sites, or contents that could have possibly resulted in those behaviors.

First we discuss an example of behavior analysis on social media and demonstrate how this behavior can be analyzed. After that, we outline the process that can be followed to analyze any behavior on social media.

#### Community Membership in Social Media

Users often join different communities in social media; the act of becoming a community member is an example of user-community behavior. Why do users join communities? In other words, what factors affect the community-joining behavior of individuals?

To analyze community-joining behavior, we can observe users who join communities and determine the factors that are common among them. Hence, we require a population of users  $U = \{u_1, u_2, \dots, u_n\}$ , a community  $C$ , and community membership information (i.e., users  $u_i \in U$  who are members of  $C$ ). The community need not be explicitly defined. For instance, one can think of individuals buying a product as a community, and people buying the product for the first time as individuals joining the community. To distinguish between users who have already joined the community and those who are now joining it, we need community memberships at two different times:  $t_1$  and  $t_2$ , with  $t_2 > t_1$ . At  $t_2$ , we determine users such as  $u$  who are currently members of the community, but were not members at  $t_1$ . These new users form the subpopulation that is analyzed for community-joining behavior.

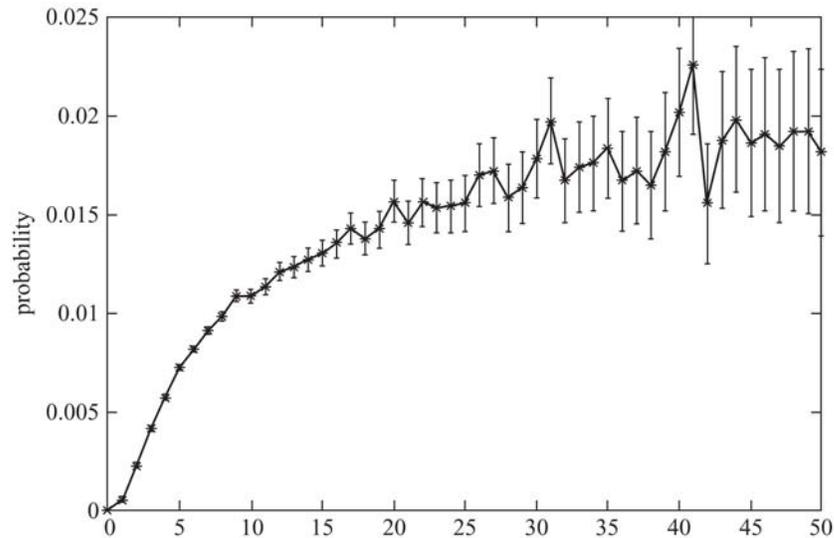


Figure 10.2: Probability of Joining a Community (with Error Bars) as a Function of the Number of Friends  $m$  Already in the Community (from [19]).

To determine factors that affect community-joining behavior, we can design hypotheses based on different factors that describe when community-joining behavior takes place. We can verify these hypotheses by using data available on social media. The factors used in the validated hypotheses describe the behavior under study most accurately.

One such hypothesis is that individuals are inclined toward an activity when their friends are engaged in the same activity. Thus, if the hypothesis is valid, a factor that plays a role in users joining a community is the number of their friends who are already members of the community. In data mining terms, this translates to using the number of friends of an individual in a community as a feature to predict whether the individual joins the community (i.e., class attribute). Figure 10.2 depicts the probability of joining a community with respect to the number of friends an individual has who are already members of the community. The probability increases as more friends are in a community, but a *diminishing returns* property is also observed, meaning that when enough friends are inside the community, more friends have no or only marginal effects on the likelihood of the individual's act of joining the community.

Diminishing  
Returns

Feature Set	Feature
Features related to the community, $C$ . (Edges between only members of the community are $E_C \subseteq E$ .)	Number of members ( $ C $ ).
	Number of individuals with a friend in $C$ (the <i>fringe</i> of $C$ ).
	Number of edges with one end in the community and the other in the fringe.
	Number of edges with both ends in the community, $ E_C $ .
	The number of open triads: $ \{(u, v, w)   (u, v) \in E_C \wedge (v, w) \in E_C \wedge (u, w) \notin E_C \wedge u \neq w\} $ .
	The number of closed triads: $ \{(u, v, w)   (u, v) \in E_C \wedge (v, w) \in E_C \wedge (u, w) \in E_C\} $ .
	The ratio of closed to open triads.
	The fraction of individuals in the fringe with at least $k$ friends in the community for $2 \leq k \leq 19$ .
	The number of posts and responses made by members of the community.
	The number of members of the community with at least one post or response.
The number of responses per post.	
Features related to an individual $u$ and her set $S$ of friends in community $C$ .	Number of friends in community ( $ S $ ).
	Number of adjacent pairs in $S$ ( $ \{(u, v)   u, v \in S \wedge (u, v) \in E_C\} $ ).
	Number of pairs in $S$ connected via a path in $E_C$ .
	Average distance between friends connected via a path in $E_C$ .
	Number of community members reachable from $S$ using edges in $E_C$ .
	Average distance from $S$ to reachable community members using edges in $E_C$ .
	The number of posts and response made by individuals in $S$ .
The number of individuals in $S$ with at least 1 post or response.	

Figure 10.3: User Community-Joining Behavior Features (from [19]).

Thus far we have defined only one feature. However, one can go beyond a single feature. Figure 10.3 lists the comprehensive features that can be used to analyze community-joining behavior.

As discussed, these features may or may not affect the joining behavior; thus, a validation procedure is required to understand their effect on the joining behavior. Which one of these features is more relevant to the joining behavior? In other words, which feature can help best determine whether individuals will join or not?

To answer this question, we can use any *feature selection* algorithm. Feature selection algorithms determine features that contribute the most to the prediction of the class attribute. Alternatively, we can use a classification algorithm, such as decision tree learning, to identify the relationship between features and the class attribute (i.e.,  $\text{joined} = \{\text{Yes}, \text{No}\}$ ). The earlier a feature is selected in the learned tree (i.e., is closer to the root of the tree), the more important to the prediction of the class attribute value.

By performing decision tree learning for a large dataset of users and the features listed in Figure 10.3, one finds that not only the number of friends inside a community but also how these friends are connected to each other affect the joining probability. In particular, the denser the subgraph of friends inside a community, the higher the likelihood of a user joining the community. Let  $S$  denote the set of friends inside community  $C$ , and let  $E_S$  denote the set of edges between these  $|S|$  friends. The maximum number of edges between these  $S$  friends is  $\binom{|S|}{2}$ . So, the edge density is  $\phi(S) = E_S / \binom{|S|}{2}$ .

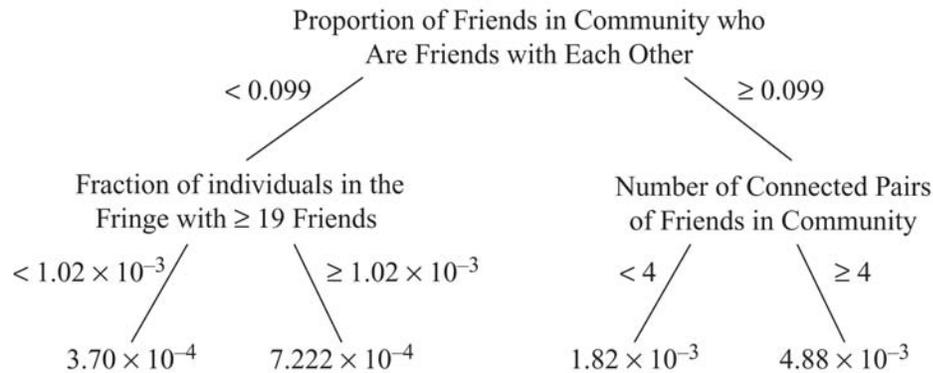


Figure 10.4: Decision Tree Learned for Community-Joining Behavior (from [19]).

One finds that the higher this density, the more likely that one is going to join a community. Figure 10.4 shows the first two levels of the decision tree learned for this task using features described in Figure 10.3. Higher level features are more discriminative in decision tree learning, and in our case, the most important feature is the density of edges for the friends subgraph inside the community.

To analyze community-joining behavior, one can design features that are likely to be related to community joining behavior. Decision tree learning can help identify which features are more predictive than others. However, how can we evaluate if these features are designed well and whether other features are not required to accurately predict joining behavior? Since classification is used to learn the relation between features and behaviors one can always use classification evaluation metrics such as accuracy to evaluate the performance of the learned model. An accurate model translates to an accurate learning of feature-behavior association.

### A Behavior Analysis Methodology

The analysis of community-joining behavior can be summarized via a four-step methodology for behavioral analysis. The same approach can be followed as a general guideline for analyzing other behaviors in social media.

Commonly, to perform behavioral analysis, one needs the following four components:

1. **An observable behavior.** The behavior that is analyzed needs to be observable. For instance, to analyze community-joining behavior, it is necessary to be able to accurately observe the joining of individuals (and possibly their joining times).
2. **Features.** One needs to construct relevant data features (covariates) that may or may not affect (or be affected by) the behavior. Anthropologists and sociologists can help design these features. The intrinsic relation between these features and the behavior should be clear from the domain expert's point of view. In community-joining behavior, we used the number of friends inside the community as one feature.
3. **Feature-Behavior Association.** This step aims to find the relationship between features and behavior, which describes how changes in features result in the behavior (or changes its intensity). We used decision tree learning to find features that are most correlated with community-joining behavior.
4. **Evaluation Strategy.** The final step evaluates the findings. This evaluation guarantees that the findings are due to the features defined and not to externalities. We use classification accuracy to verify the quality of features in community-joining behavior. Various evaluation techniques can be used, such as randomization tests discussed in Chapter 8. In randomization tests, we measure a phenomenon in a dataset and then randomly generate subsamples from the dataset in which the phenomenon is guaranteed to be removed. We assume the phenomenon has happened when the measurements on the subsamples are different from the ones on the original dataset. Another approach is to use *causality testing* methods. Causality testing methods measure how a feature can affect a phenomenon. A well-known causality detection technique is called *granger causality* due to Clive W. J. Granger, the Nobel laureate in economics.

Causality  
Testing

Granger  
Causality

**Definition 10.1. Granger Causality.** Assume we are given two temporal variables  $X = \{X_1, X_2, \dots, X_t, X_{t+1}, \dots\}$  and  $Y = \{Y_1, Y_2, \dots, Y_t, Y_{t+1}, \dots\}$ . Variable  $X$  “Granger causes” variable  $Y$  when historical values of  $X$  can help better predict  $Y$  than just using the historical values of  $Y$ .

Consider a linear regression model outlined in Chapter 5. We can predict  $Y_{t+1}$  by using either  $Y_1, \dots, Y_t$  or a combination of  $X_1, \dots, X_t$  and  $Y_1, \dots, Y_t$ .

$$Y_{t+1} = \sum_{i=1}^t a_i Y_i + \epsilon_1, \quad (10.1)$$

$$Y_{t+1} = \sum_{i=1}^t a_i Y_i + \sum_{i=1}^t b_i X_i + \epsilon_2, \quad (10.2)$$

where  $\epsilon_1$  and  $\epsilon_2$  are the regression model errors. Now, if  $\epsilon_2 < \epsilon_1$ , it indicates that using  $X$  helps reduce the error. In this case,  $X$  Granger causes  $Y$ .

## 10.1.2 Individual Behavior Modeling

Similar to network models, models of individual behavior can help concretely describe why specific individual behaviors are observed in social media. In addition, they allow for controlled experiments and simulations that can help study individuals in social media.

As with other modeling approaches (see Chapter 4), in behavior modeling, one must make a set of assumptions. Behavior modeling can be performed via a variety of techniques, including those from economics, game theory, or network science. We discussed some of these techniques in earlier chapters. We review them briefly here, and refer interested readers to the respective chapters for more details.

- **Threshold models (Chapter 8).** When a behavior diffuses in a network, such as the behavior of individuals buying a product and referring it to others, one can use threshold models. In threshold models, the parameters that need to be learned are the node activation threshold  $\theta_i$  and the influence probabilities  $w_{ij}$ . Consider the following methodology for learning these values. Consider a merchandise store where the store knows the connections between individuals and their transaction history (e.g., the items that they have bought). Then,  $w_{ij}$  can be defined as the

fraction of times user  $i$  buys a product and  
user  $j$  buys the same product **soon** after that

The definition of “soon” requires clarification and can be set based on a site’s preference and the average time between friends buying the same product. Similarly,  $\theta_i$  can be estimated by taking into account the average number of friends who need to buy a product before user  $i$  decides to buy it. Of course, this is only true when the products bought by user  $i$  are also bought by her friends. When this is not the case, methods from collaborative filtering (see Chapter 9) can be used to find out the average number of **similar** items that are bought by user  $i$ ’s friends before user  $i$  decides to buy a product.

- **Cascade Models (Chapter 7).** Cascade models are examples of scenarios where an innovation, product, or information cascades through a network. The discussion with respect to cascade models is similar, to the threshold models with the exception that cascade models are sender-centric. That is, the sender decides to activate the receiver, whereas threshold models are receiver-centric, in which receivers get activated by multiple senders. Therefore, the computation of the ICM parameters needs to be done from the sender’s point of view in cascade models. Note that both threshold and cascade models are examples of individual behavior modeling.

### 10.1.3 Individual Behavior Prediction

As discussed previously, most behaviors result in newly formed links in social media. It can be a link to a user, as in befriending behavior; a link to an entity, as in buying behavior; or a link to a community, as in joining behavior. Hence, one can formulate many of these behaviors as a link prediction problem. Next, we discuss link prediction in social media.

#### Link Prediction

Link prediction assumes a graph  $G(V, E)$ . Let  $e(u, v) \in E$  represent an interaction (edge) between nodes  $u$  and  $v$ , and let  $t(e)$  denote the time of the interaction. Let  $G[t_1, t_2]$  represent the subgraph of  $G$  such that all edges are created between  $t_1$  and  $t_2$  (i.e., for all edges  $e$  in this subgraph,  $t_1 < t(e) < t_2$ ). Now given four time stamps  $t_1 < t'_1 < t_2 < t'_2$ , a link prediction algorithm is given the subgraph  $G[t_1, t'_1]$  (training interval) and is expected to predict edges in  $G[t_2, t'_2]$  (testing interval). Note that, just

like new edges, new nodes can be introduced in social networks; therefore,  $G[t_2, t'_2]$  may contain nodes not present in  $G[t_1, t'_1]$ . Hence, a link prediction algorithm is generally constrained to predict edges only for pairs of nodes that are present during the training period. One can add extra constraints such as predicting links only for nodes that are incident to at least  $k$  edges (i.e., have degree greater or equal to  $k$ ) during both testing and training intervals.

Let  $G(V_{train}, E_{train})$  be our training graph. Then, a link prediction algorithm generates a sorted list of most probable edges in  $V_{train} \times V_{train} - E_{train}$ . The first edge in this list is the one the algorithm considers the most likely to soon appear in the graph. The link prediction algorithm assigns a score  $\sigma(x, y)$  to every edge  $e(x, y)$  in  $V_{train} \times V_{train} - E_{train}$ . Edges sorted by this value in decreasing order will create our ranked list of predictions.  $\sigma(x, y)$  can be predicted based on different techniques. Note that any similarity measure between two nodes can be used for link prediction; therefore, methods discussed in Chapter 3 are of practical use here. We outline some of the most well-established techniques for computing  $\sigma(x, y)$  here.

### Node Neighborhood-Based Methods

The following methods take advantage of neighborhood information to compute the similarity between two nodes.

- **Common Neighbors.** In this method, one assumes that the more common neighbors that two nodes share, the more similar they are. Let  $N(x)$  denote the set of neighbors of node  $x$ . This method is formulated as

$$\sigma(x, y) = |N(x) \cap N(y)|. \quad (10.3)$$

- **Jaccard Similarity.** This commonly used measure calculates the likelihood of a node that is a neighbor of either  $x$  or  $y$  to be a common neighbor. It can be formulated as the number of common neighbors divided by the total number of neighbors of either  $x$  or  $y$ :

$$\sigma(x, y) = \frac{|N(x) \cap N(y)|}{|N(x) \cup N(y)|}. \quad (10.4)$$

- **Adamic and Adar Measure.** A similar measure to Jaccard, this measure was introduced by Lada Adamic and Eytan Adar [2003]. The

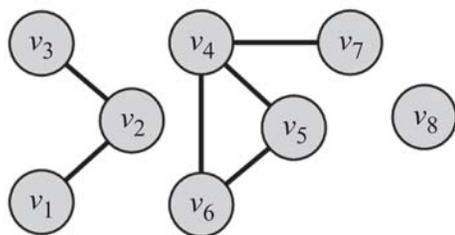


Figure 10.5: Neighborhood-Based Link Prediction Example.

intuition behind it measure is that if two individuals share a neighbor and that neighbor is a *rare* neighbor, it should have a higher impact on their similarity. For instance, we can define the rareness of a node based on its degree (i.e., the smaller the node's degree, the higher its rareness). The original version of the measure is defined based on webpage features. A modified version based on neighborhood information is

$$\sigma(x, y) = \sum_{z \in N(x) \cap N(y)} \frac{1}{\log |N(z)|}. \quad (10.5)$$

- **Preferential Attachment.** In the preferential attachment model discussed in Chapter 4, one assumes that nodes of higher degree have a higher chance of getting connected to incoming nodes. Therefore, in terms of connection probability, higher degree nodes are similar. The preferential attachment measure is defined to capture this similarity:

$$\sigma(x, y) = |N(x)| \cdot |N(y)|. \quad (10.6)$$

**Example 10.1.** For the graph depicted in Figure 10.5, the similarity between nodes 5 and 7 based on different neighborhood-based techniques is

$$\text{(Common Neighbor)} \sigma(5, 7) = |\{4, 6\} \cap \{4\}| = 1 \quad (10.7)$$

$$\text{(Jaccard)} \sigma(5, 7) = \frac{|\{4, 6\} \cap \{4\}|}{|\{4, 6\} \cup \{4\}|} = \frac{1}{2} \quad (10.8)$$

$$\text{(Adamic and Adar)} \sigma(5, 7) = \frac{1}{\log |\{5, 6, 7\}|} = \frac{1}{\log 3} \quad (10.9)$$

$$\text{(Preferential Attachment)} \sigma(5, 7) = |\{4\}| \cdot |\{4, 6\}| = 1 \times 2 = 2 \quad (10.10)$$

Table 10.1: A Comparison between Link Prediction Methods

	First edge	Second edge	Third edge
<i>Common Neighbors</i>	$\sigma(6,7) = 1$	$\sigma(1,3) = 1$	$\sigma(5,7) = 1$
<i>Jaccard Similarity</i>	$\sigma(1,3) = 1$	$\sigma(6,7) = 1/2$	$\sigma(5,7) = 1/2$
<i>Adamic and Adar</i>	$\sigma(1,3) = 1/\log 2$	$\sigma(6,7) = 1/\log 3$	$\sigma(5,7) = 1/\log 3$
<i>Preferential Attachment</i>	$\sigma(2,4) = 6$	$\sigma(2,5) = 4$	$\sigma(2,6) = 4$

In Figure 10.5, there are eight nodes; therefore, we can have a maximum of  $\binom{8}{2} = 28$  edges. We already have six edges in the graph; hence, there are  $28 - 6 = 22$  other edges that are not in the graph. For all these edges, we can compute the similarity between their endpoints using the aforementioned neighborhood-based techniques and identify the top three most likely edges that are going to appear in the graph based on each technique. Table 10.1 shows the top three edges based on each technique and the corresponding values for each edge. As shown in this table, different methods predict different edges to be most important; therefore, the method of choice depends on the application.

### Methods Based on Paths between Nodes

Similarity between nodes can simply be computed from the shortest path distance between them. The closer the nodes are, the higher their similarity. This similarity measure can be extended by considering multiple paths between nodes and their neighbors. The following measures can be used to calculate similarity.

- **Katz measure.** Similar to the Katz centrality defined in Chapter 3, one can define the similarity between nodes  $x$  and  $y$  as

$$\sigma(x, y) = \sum_{l=1}^{\infty} \beta^l |\text{paths}_{x,y}^{<l>}|, \quad (10.11)$$

where  $|\text{paths}_{x,y}^{<l>}|$  denotes the number of paths of length  $l$  between  $x$  and  $y$ .  $\beta$  is a constant that exponentially damps longer paths. Note that a very small  $\beta$  results in a common neighbor measure (see Exercises). Similar to our finding in Chapter 3, one can find the Katz similarity measure in a closed form by  $(I - \beta A)^{-1} - I$ . The Katz measure can also

be weighted or unweighted. In the unweighted format,  $|paths_{x,y}^{<1>}| = 1$  if there is an edge between  $x$  and  $y$ . The weighted version is more suitable for multigraphs, where multiple edges can exist between the same pair of nodes. For example, consider two authors  $x$  and  $y$  who have collaborated  $c$  times. In this case,  $|paths_{x,y}^{<1>}| = c$ .

- **Hitting and Commute Time.** Consider a random walk that starts at node  $x$  and moves to adjacent nodes uniformly. Hitting time  $H_{x,y}$  is the expected number of random walk steps needed to reach  $y$  starting from  $x$ . This is a distance measure. In fact, a smaller hitting time implies a higher similarity; therefore, a negation can turn it into a similarity measure:

$$\sigma(x, y) = -H_{x,y}. \quad (10.12)$$

Note that if node  $y$  is highly connected to other nodes in the network (i.e., has a high stationary probability  $\pi_y$ ), then a random walk starting from any  $x$  likely ends up visiting  $y$  early. Hence, all hitting times to  $y$  are very short, and all nodes become similar to  $y$ . To account for this, one can normalize hitting time by multiplying it with the stationary probability  $\pi_y$ :

$$\sigma(x, y) = -H_{x,y}\pi_y. \quad (10.13)$$

Hitting time is not symmetric, and in general,  $H_{x,y} \neq H_{y,x}$ . Thus, one can introduce the *commute* time to mitigate this issue:

$$\sigma(x, y) = -(H_{x,y} + H_{y,x}). \quad (10.14)$$

Similarly, commute time can also be normalized,

$$\sigma(x, y) = -(H_{x,y}\pi_y + H_{y,x}\pi_x). \quad (10.15)$$

- **Rooted PageRank.** A modified version of the PageRank algorithm can be used to measure similarity between two nodes  $x$  and  $y$ . In rooted PageRank, we measure the stationary probability of  $y$ :  $\pi_y$  given the condition that during each random walk step, we jump to  $x$  with probability  $\alpha$  or a random neighbor with probability  $1 - \alpha$ . The matrix format discussed in Chapter 3 can be used to solve this problem.

- **SimRank.** One can define similarity between two nodes recursively based on the similarity between their neighbors. In other words, similar nodes have similar neighbors. SimRank performs the following:

$$\sigma(x, y) = \gamma \cdot \frac{\sum_{x' \in N(x)} \sum_{y' \in N(y)} \sigma(x', y')}{|N(x)||N(y)|}, \quad (10.16)$$

where  $\gamma$  is some value in range  $[0, 1]$ . We set  $\sigma(x, x) = 1$ , and by finding the fixed point of this equation, we can find the similarity between node  $x$  and node  $y$

After one of the aforementioned measures is selected, a list of the top most similar pairs of nodes are selected. These pairs of nodes denote edges predicted to be the most likely to soon appear in the network. Performance (precision, recall, or accuracy) can be evaluated using the testing graph and by comparing the number of the testing graph's edges that the link prediction algorithm successfully reveals. Note that the performance is usually very low, since many edges are created due to reasons not solely available in a social network graph. So, a common baseline is to compare the performance with random edge predictors and report the factor improvements over random prediction.

## 10.2 Collective Behavior

Collective behavior, first defined by sociologist Robert Park, refers to a population of individuals behaving in a similar way. This similar behavior can be planned and coordinated, but is often spontaneous and unplanned. For instance, individuals stand in line for a new product release, rush into stores for a sale event, and post messages online to support their cause or show their support for an individual. These events, though formed by independent individuals, are observed as a collective behavior by outsiders.

### 10.2.1 Collective Behavior Analysis

Collective behavior analysis is often performed by analyzing individuals performing the behavior. In other words, one can divide collective behavior into many individual behaviors and analyze them independently.

The result, however, when all these analyses are put together would be an *expected* behavior for a large population. The user migration behavior we discuss in this section is an example of this type of analysis of collective behavior.

One can also analyze the population as a whole. In this case, an individual's opinion or behavior is rarely important. In general, the approach is the same as analyzing an individual, with the difference that the content and links are now considered for a large community. For instance, if we are analyzing 1,000 nodes, one can combine these nodes and edges into one hyper-node, where the hyper-node is connected to all other nodes in the graph to which its members are connected and has an internal structure (subgraph) that details the interaction among its members. This approach is unpopular for **analyzing** collective behavior because it does not consider specific individuals and at times, interactions within the population. Interested readers can refer to the bibliographic notes for further references that use this approach to analyze collective behavior. On the contrary, this approach is often considered when **predicting** collective behavior, which is discussed later in this chapter.

### User Migration in Social Media

Users often migrate from one site to another for different reasons. The main rationale behind it is that users have to select some sites over others due to their limited time and resources. Moreover, social media's networking often dictates that one cannot freely choose a site to join or stay. An individual's decision is heavily influenced by his or her friends, and vice versa. Sites are often interested in keeping their users, because they are valuable assets that help contribute to their growth and generate revenue by increased traffic. There are two types of migration that take place in social media sites: *site migration* and *attention migration*.

1. **Site Migration.** For any user who is a member of two sites  $s_1$  and  $s_2$  at time  $t_i$ , and is only a member of  $s_2$  at time  $t_j > t_i$ , then the user is said to have migrated from site  $s_1$  to site  $s_2$ .
2. **Attention Migration.** For any user who is a member of two sites  $s_1$  and  $s_2$  and is active at both at time  $t_i$ , if the user becomes inactive on  $s_1$  and remains active on  $s_2$  at time  $t_j > t_i$ , then the user's attention is said to have migrated away from site  $s_1$  and toward site  $s_2$ .

Activity (or inactivity) of a user can be determined by observing the user's actions performed on the site. For instance, we can consider a user active in interval  $[t_i, t_i + \delta]$ , if the user has performed at least one action on the site during this interval. Otherwise, the user is considered inactive.

The interval  $\delta$  could be measured at different granularity, such as *days*, *weeks*, *months*, and *years*. It is common to set  $\delta = 1$  month. To analyze the migration of populations across sites, we can analyze migrations of individuals and then measure the rate at which the population of these individuals is migrating across sites. Since this method analyzes migrations at the individual level, we can use the methodology outlined in Section 10.1.1 for individual behavior analysis as follows.

### The Observable Behavior

Site migration is rarely observed since users often abandon their accounts rather than closing them. A more observable behavior is attention migration, which is clearly observable on most social media sites. Moreover, when a user commits site migration, it is often too late to perform preventive measures. However, when attention migration is detected, it is still possible to take actions to retain the user or expedite his or her attention migration to guarantee site migration. Thus, we focus on individuals whose attention has migrated.

To observe attention migrations, several steps need to be taken. First, users are required to be identified on multiple networks so that their activity on multiple sites can be monitored simultaneously. For instance, username `huan.liu1` on Facebook is username `liuhuan` on Twitter. This identification can be done by collecting information from sites where individuals list their multiple identities on social media sites. On social networking sites such as Google+ or Facebook, this happens regularly. The second step is collecting multiple snapshots of social media sites. At least two snapshots are required to observe migrations. After these two steps, we can observe whether attention migrations have taken place or not. In other words, we can observe if users have become inactive on one of the sites over time. Figure 10.6 depicts these migrations for some well-known social media sites. In this figure, each radar chart shows migrations *from* a site to multiple sites. Each target site is shown as a vertex, and the longer the spokes toward that site, the larger the migrating population to it.

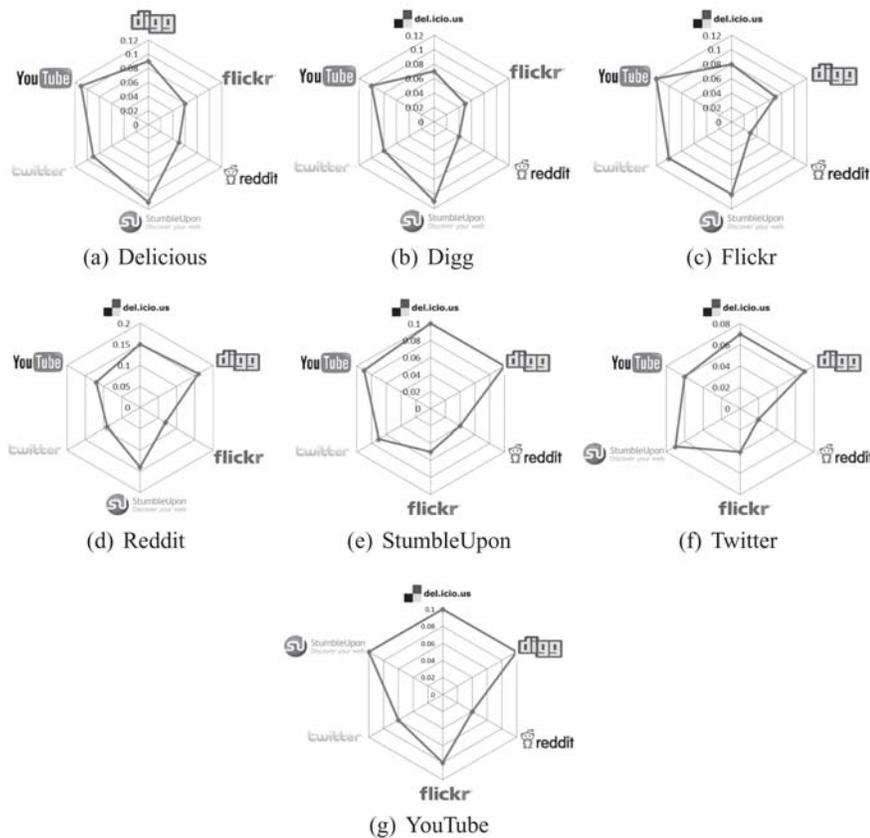


Figure 10.6: Pairwise Attention Migration among Social Media Sites.

### Features

Three general features can be considered for user migration: (1) *user activity* on one site, (2) *user network size*, and (3) *user rank*. User activity is important, because we can conjecture that a more active user on one site is less likely to migrate. User network size is important, because a user with more social ties (i.e., friends) in a social network is less likely to move. Finally, user rank is important. The rank is the value of a user as perceived by others. A user with high status in a network is less likely to move to a new one where he or she must spend more time getting established.

User activity can be measured differently for different sites. On Twitter, it can be the number of tweets posted by the user; on Flickr, the number photos uploaded by the user; and on YouTube, the number of videos the

user has uploaded. One can normalize this value by its maximum in the site (e.g., the maximum number of videos any user has uploaded) to get an activity measure in the range  $[0,1]$ . If a user is allowed to have multiple activities on a site, as in posting comments and liking videos, then a linear combination of these measures can be used to describe user activity on a site.

User network size can be easily measured by taking the number of friends a user has on the site. It is common for social media sites to facilitate the addition of friends. The number of friends can be normalized in the range  $[0,1]$  by the maximum number of friends one can have on the site.

Finally, user rank is how important a user is on the site. Some sites explicitly provide their users' prestige rank list (e.g., top 100 bloggers), whereas for others, one needs to approximate a user's rank. One way of approximating it is to count the number of citations (in-links) an individual is receiving from others. A practical technique is to perform this via web search engines. For instance, user test on StumbleUpon has <http://test.stumbleupon.com> as his profile page. A Google search for `link:http://test.stumbleupon.com` provides us with the number of in-links to the profile on StumbleUpon and can be considered as a ranking measure for user test.

These three features are correlated with the site attention migration behavior and one expects changes in them when migrations happen.

### **Feature-Behavior Association**

Given two snapshots of a network, we know if users migrated or not. We can also compute the values for the aforementioned features. Hence, we can determine the correlation between features and migration behavior.

Let vector  $Y \in \mathbb{R}^n$  indicate whether any of our  $n$  users have migrated or not. Let  $X_t \in \mathbb{R}^{3 \times n}$  be the features collected (activity, friends, rank) for any one of these users at time stamp  $t$ . Then, the correlation between features  $X_t$  and labels  $Y$  can be computed via logistic regression. How can we verify that this correlation is not random? Next, we discuss how we verify that this correlation is statistically significant.

## Evaluation Strategy

To verify if the correlation between features and the migration behavior is not random, we can construct a random set of migrating users and compute  $X_{\text{Random}}$  and  $Y_{\text{Random}}$  for them as well. This can be obtained by shuffling the rows of the original  $X_t$  and  $Y$ . Then, we perform logistic regression on these new variables. This approach is very similar to the shuffle test presented in Chapter 8. The idea is that if some behavior creates a change in features, then other random behaviors should not create that drastic a change. So, the observed correlation between features and the behavior should be significantly different in both cases. The correlation can be described in terms of logistic regression coefficients, and the significance can be measured via any significance testing methodology. For instance, we can employ the  $\chi^2$ -statistic,  $\chi^2$ -statistic

$$\chi^2 = \sum_{i=1}^n \frac{(A_i - R_i)^2}{R_i}, \quad (10.17)$$

where  $n$  is the number of logistic regression coefficients,  $A_i$ 's are the coefficients determined using the original dataset, and  $R_i$ 's are the coefficients obtained from the random dataset.

### 10.2.2 Collective Behavior Modeling

Consider a hypothetical model that can simulate voters who cast ballots in elections. This effective model can help predict an election's turnout rate as an outcome of the collective behavior of voting and help governments prepare logistics accordingly. This is an example of collective behavior modeling, which improves our understanding of the collective behaviors that take place by providing concrete explanations.

Collective behavior can be conveniently modeled using some of the techniques discussed in Chapter 4, "Network Models". Similar to collective behavior, in network models, we express models in terms of characteristics observable in the population. For instance, when a power-law degree distribution is required, the preferential attachment model is preferred, and when the small average shortest path is desired, the small-world model is the method of choice. In network models, node properties rarely play a role; therefore, they are reasonable for modeling collective behavior.

### 10.2.3 Collective Behavior Prediction

Collective behavior can be predicted using methods we discussed in Chapters 7 and 8. For instance, epidemics can predict the effect of a disease on a population and the behavior that the population will exhibit over time. Similarly, implicit influence models such as the LIM model discussed in Chapter 8 can estimate the influence of individuals based on collective behavior attributes, such as the size of the population adopting an innovation at any time.

As noted earlier, collective behavior can be analyzed either in terms of individuals performing the collective behavior or based on the population as a whole. When predicting collective behavior, it is more common to consider the population as a whole and aim to predict some phenomenon. This simplifies the challenges and reduces the computation dramatically, since the number of individuals who perform a collective behavior is often large and analyzing them one at a time is cumbersome.

In general, when predicting collective behavior, we are interested in predicting the intensity of a phenomenon, which is due to the collective behavior of the population (e.g., how many of them will vote?) To perform this prediction, we utilize a data mining approach where features that describe the population well are used to predict a response variable (i.e., the intensity of the phenomenon). A training-testing framework or correlation analysis is used to determine the generalization and the accuracy of the predictions. We discuss this collective behavior prediction strategy through the following example. This example demonstrates how the collective behavior of individuals on social media can be utilized to predict real-world outcomes.

#### Predicting Box Office Revenue for Movies

Can we predict opening-weekend revenue for a movie from its prerelease chatter among fans? This tempting goal of *predicting the future* has been around for many years. The goal is to predict the collective behavior of watching a movie by a large population, which in turn determines the revenue for the movie. One can design a methodology to predict box office revenue for movies that uses Twitter and the aforementioned collective behavior prediction strategy. To summarize, the strategy is as follows:

1. Set the target variable that is being predicted. In this case, it is the

revenue that a movie produces. Note that the revenue is the direct result of the collective behavior of going to the theater to watch the movie.

2. Determine the features in the population that may affect the target variable.
3. Predict the target variable using a supervised learning approach, utilizing the features determined in step 2.
4. Measure performance using supervised learning evaluation.

One can use the population that is discussing the movie on Twitter before its release to predict its opening-weekend revenue. The target variable is the amount of revenue. In fact, utilizing only eight features, one can predict the revenue with high accuracy. These features are the average hourly number of tweets related to the movie for each of the seven days prior to the movie opening (seven features) and the number of opening theaters for the movie (one feature). Using only these eight features, training data for some movies (their seven-day tweet rates, their number of opening theaters, and their revenue), and a linear regression model, one can predict the movie opening-weekend revenue with high correlation. It has been shown by researchers (see Bibliographic Notes) that the predictions using this approach are closer to reality than that of the Hollywood Stock Exchange (HSX), which is the gold standard for predicting revenues for movies.

This simple model for predicting movie revenue can be easily extended to other domains. For instance, assume we are planning to predict another collective behavior outcome, such as the number of individuals who aim to buy a product. In this case, the target variable  $y$  is the number of individuals who will buy the product. Similar to tweet rate, we require some feature  $A$  that denotes the attention the product is receiving. We also need to model the publicity of the product  $P$ . In our example, this was the number of theaters for the movie; for a product, it could represent the number of stores that sell it. A simple linear regression model can help learn the relation between these features and the target variable:

$$y = w_1A + w_2P + \epsilon, \quad (10.18)$$

where  $\epsilon$  is the regression error. Similar to our movie example, one attempts to extract the values for  $A$  and  $P$  from social media.

## 10.3 Summary

Individuals exhibit different behaviors in social media, which can be categorized into individual and collective behavior. Individual behavior is the behavior that an individual targets toward (1) another individual (*individual-individual behavior*), (2) an entity (*individual-entity behavior*), or (3) a community (*individual-community behavior*). We discussed how to analyze and predict individual behavior. To analyze individual behavior, there is a four-step procedure, outlined as a guideline. First, the behavior observed should be clearly observable on social media. Second, one needs to design meaningful features that are correlated with the behavior taking place in social media. The third step aims to find correlations and relationships between features and the behavior. The final step is to verify these relationships that are found. We discussed community joining as an example of individual behavior. Modeling individual behavior can be performed via cascade or threshold models. Behaviors commonly result in interactions in the form of links; therefore, link prediction techniques are highly efficient in predicting behavior. We discussed neighborhood-based and path-based techniques for link prediction.

Collective behavior is when a group of individuals with or without coordination act in an aligned manner. Collective behavior analysis is either done via individual behavior analysis and then averaged or analyzed collectively. When analyzed collectively, one commonly looks at the general patterns of the population. We discussed user migrations in social media as an example of collective behavior analysis. Modeling collective behavior can be performed via network models, and prediction is possible by using population properties to predict an outcome. Predicting movie box-office revenues was given as an example, which uses population properties such as the rate at which individuals are tweeting to demonstrate the effectiveness of this approach.

It is important to evaluate behavior analytics findings to ensure that these findings are not due to externalities. We discussed causality testing, randomization tests, and supervised learning evaluation techniques for evaluating behavior analytics findings. However, depending on the context, researchers may need to devise other informative techniques to ensure the validity of the outcomes.

## 10.4 Bibliographic Notes

In addition to methods discussed in this chapter, game theory and theories from economics can be used to analyze human behavior [78]. Community-joining behavior analysis was first introduced by Backstrom et al. [19]. The approach discussed in this chapter is a brief summary of their approach for analyzing community-joining behavior. Among other individual behaviors, tie formation is analyzed in detail. In [290], the authors analyze tie formation behavior on Facebook and investigate how visual cues influence individuals with no prior interaction to form ties. The features used are gender (i.e., male or female), and visual conditions (attractive, nonattractive, and no photo). Their analyses show that individuals have a tendency to connect to attractive opposite-sex individuals when no other information is available. Analyzing individual information-sharing behavior helps understand how individuals disseminate information on social media. Gundecha et al. [114] analyze how the information-sharing behavior of individuals results in vulnerabilities and how one can exploit such vulnerabilities to secure user privacy on a social networking site. Finally, most social media mining research is dedicated to analyzing a single site; however, users are often members of different sites and hence, current studies need to be generalized to cover multiple sites. Zafarani and Liu [311, 312] were the first to design methods that help connect user identities across social media sites using behavioral modeling. A study of user tagging behavior across sites is available in [291].

General surveys on link prediction can be found in [4, 172, 8, 179]. Individual behavior prediction is an active area of research. Location prediction is an active area of individual behavior analysis that has been widely studied over a long period in the realm of mobile computing. Researchers analyze human mobility patterns to improve location prediction services, thereby exploiting their potential power on various applications such as mobile marketing [29, 27], traffic planning [33, 70], and even disaster relief [98, 94, 105, 95, 289, 26, 160]. Other general references can be found in [20, 200, 263, 282, 247, 96, 97].

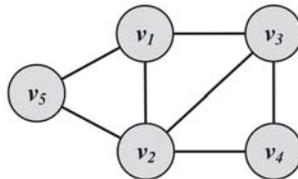
Kumar et al. [159] first analyzed migration in social media. Other collective behavior analyses can be found in Leskovec et al. [170]. The movie revenue prediction was first discussed by Asur and Huberman [18]. Another example of collective behavior prediction can be found in the work of O'Connor et al. [221], which proposed using Twitter data for opinion

polls. Their results are highly correlated with Gallup opinion polls for presidential job approval. In [1], the authors analyzed collective social media data and show that by carefully selecting data from social media, it is possible to use social media as a lens to analyze and even predict real-world events.

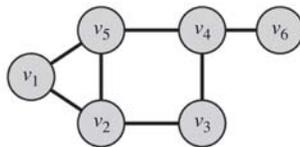
## 10.5 Exercises

### Individual Behavior

- Name five real-world behaviors that are commonly difficult to observe in social media (e.g., your daily schedule or where you eat lunch are rarely available in social media).
  - Select one behavior that is most likely to leave traces online. Can you think of a methodology for identifying that behavior using these traces?
2. Consider the “commenting under a blogpost” behavior in social media. Follow the four steps of behavior analysis to analyze this behavior.
3. We emphasized selecting meaningful features for analyzing a behavior. Discuss a methodology to verify if the selected features carry enough information with respect to the behavior being analyzed.
4. Correlation does not imply causality. Discuss how this fact relates to most of the datasets discussed in this chapter being temporal.
5. Using a neighborhood-based link prediction method compute the top two most likely edges for the following figure.



6. Compute the most likely edge for the following figure for each path-based link prediction technique.



7. In a link prediction problem, show that for small  $\beta$ , the *Katz* similarity measure ( $\sigma(u, v) = \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\text{path}_{u,v}^{<\ell>}|$ ) becomes *Common neighbors* ( $\sigma(u, v) = |N(u) \cap N(v)|$ ).
8. Provide the matrix format for rooted PageRank and SimRank techniques.

### **Collective Behavior**

9. Recent research has shown that social media can help replicate survey results for elections and ultimately predict presidential election outcomes. Discuss what possible features can help predict a presidential election.