Chapter 8

Influence and Homophily

Social forces connect individuals in different ways. When individuals get connected, one can observe distinguishable patterns in their connectivity networks. One such pattern is *assortativity*, also known as *social similarity*. In networks with assortativity, similar nodes are connected to one another more often than dissimilar nodes. For instance, in social networks, a high similarity between friends is observed. This similarity is exhibited by similar behavior, similar interests, similar activities, and shared attributes such as language, among others. In other words, friendship networks are *assortative*. Investigating assortativity patterns that individuals exhibit on social media helps one better understand user interactions. Assortativity is the most commonly observed pattern among linked individuals. This chapter discusses assortativity along with principal factors that result in assortative networks.

Assortativity

Many social forces induce assortative networks. Three common forces are *influence*, *homophily*, and *confounding*. Influence is the process by which an individual (the influential) affects another individual such that the influenced individual becomes more similar to the influential figure. Homophily is observed in already similar individuals. It is realized when similar individuals become friends due to their high similarity. Confounding is the environment's effect on making individuals similar. For instance, individuals who live in Russia speak Russian fluently because of the environment and are therefore similar in language. The confounding force is an external factor that is independent of inter-individual interactions and is therefore not discussed further.

Influence, Homophily, and Confounding

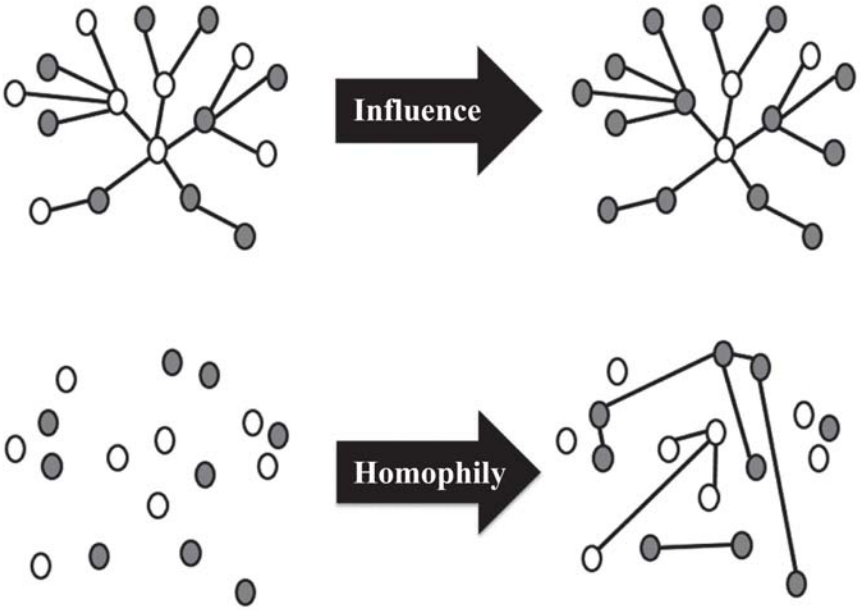


Figure 8.1: Influence and Homophily.

Note that both influence and homophily social forces give rise to assortative networks. After either of them affects a network, the network exhibits more similar nodes; however, when "friends become similar," we denote that as influence, and when "similar individuals become friends," we call it homophily. Figure 8.1 depicts how both influence and homophily affect social networks.

In particular, when discussing influence and homophily in social media, we are interested in asking the following questions:

- How can we measure influence or homophily?
- How can we model influence or homophily?
- How can we distinguish between the two?

Because both processes result in assortative networks, we can quantify their effect on the network by measuring the assortativity of the network.

8.1 Measuring Assortativity

Measuring assortativity helps quantify how much influence and homophily, among other factors, have affected a social network. Assortativity can be quantified by measuring how similar connected nodes are to one another. Figure 8.2 depicts the friendship network in a U.S. high school in 1994.[1] In the figure, races are represented with different colors: whites are white, blacks are gray, Hispanics are light gray, and others are black. As we observe, there is a high assortativity between individuals of the same race, particularly among whites and among blacks. Hispanics have a high tendency to become friends with whites.

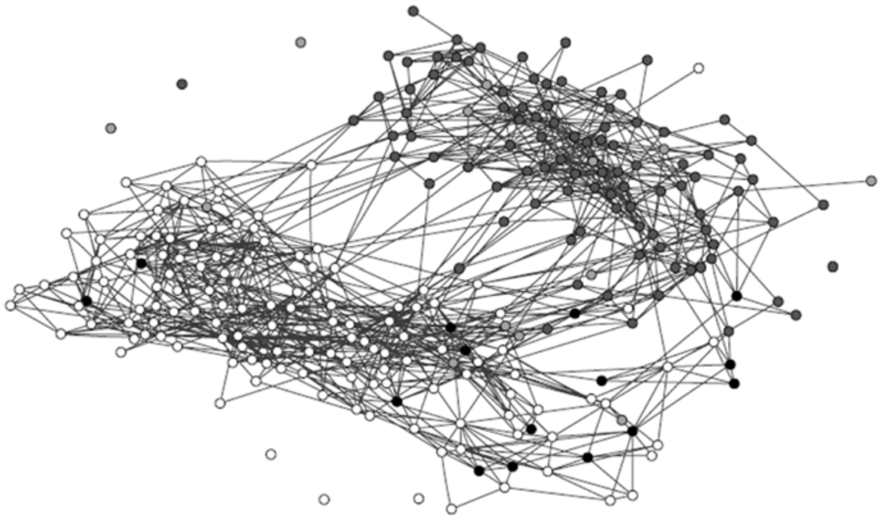


Figure 8.2: A U.S. High School Friendship Network in 1994 between Races. Eighty percent of the links exist between members of the same race (from [62]).

To measure assortativity, we measure the number of edges that fall in between the nodes of the same race. This technique works for nominal attributes, such as race, but does not work for ordinal ones such as age. Consider a network where individuals are friends with people of different ages. Unlike races, individuals are more likely to be friends with others close in age, but not necessarily with ones of the exact same age. Hence,

[1] From ADD health data: http://www.cpc.unc.edu/projects/addhealth.

we discuss two techniques: one for nominal attributes and one for ordinal attributes.

8.1.1 Measuring Assortativity for Nominal Attributes

Consider a scenario where we have nominal attributes assigned to nodes. As in our example, this attribute could be race or nationality, gender, or the like. One simple technique to measure assortativity is to consider the number of edges that are between nodes of the same type. Let $t(v_i)$ denote the type of node v_i. In an undirected graph[2], $G(V, E)$, with adjacency matrix A, this measure can be computed as follows,

$$\frac{1}{m} \sum_{(v_i,v_j)\in E} \delta(\ t(v_i), t(v_j)\ ) = \frac{1}{2m} \sum_{ij} A_{ij}\, \delta(\ t(v_i), t(v_j)\ ), \tag{8.1}$$

where m is the number of edges in the graph, $\frac{1}{m}$ is applied for normalization, and the factor $\frac{1}{2}$ is added because G is undirected. $\delta(.,.)$ is the Kronecker delta function:

$$\delta(x, y) = \begin{cases} 0, & \text{if } x \neq y; \\ 1, & \text{if } x = y. \end{cases} \tag{8.2}$$

This measure has its limitations. Consider a school of Hispanic students. Obviously, all connections will be between Hispanics, and assortativity value 1 is not a significant finding. However, consider a school where half the population is white and half the population is Hispanic. It is statistically expected that 50% of the connections will be between members of different race. If connections in this school were only between whites and Hispanics and not within groups, then our observation is significant. To account for this limitation, we can employ a common technique where we measure the *assortativity significance* by subtracting the measured assortativity by the statistically expected assortativity. The higher this value, the more significant the assortativity observed.

Assortativity Significance

Consider a graph $G(V, E)$, $|E| = m$, where the degrees are known beforehand (how many friends an individual has), but the edges are not. Consider two nodes v_i and v_j, with degrees d_i and d_j, respectively. What is the expected number of edges between these two nodes? Consider node

[2] The directed case is left to the reader.

v_i. For any edge going out of v_i randomly, the probability of this edge getting connected to node v_j is $\frac{d_j}{\sum_i d_i} = \frac{d_j}{2m}$. Since the degree for v_i is d_i, we have d_i such edges; hence, the expected number of edges between v_i and v_j is $\frac{d_i d_j}{2m}$. Now, the expected number of edges between v_i and v_j that are of the same type is $\frac{d_i d_j}{2m}\delta(\, t(v_i), t(v_j)\,)$ and the expected number of edges of the same type in the whole graph is

$$\frac{1}{m}\sum_{(v_i,v_j)\in E}\frac{d_i d_j}{2m}\,\delta(\, t(v_i), t(v_j)\,) = \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\,\delta(\, t(v_i), t(v_j)\,). \tag{8.3}$$

We are interested in computing the distance between the assortativity observed and the expected assortativity:

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{ij} A_{ij}\,\delta(\, t(v_i), t(v_j)\,) - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\,\delta(\, t(v_i), t(v_j)\,) \qquad (8.4)\\ &= \frac{1}{2m}\sum_{ij}(\, A_{ij} - \frac{d_i d_j}{2m}\,)\,\delta(\, t(v_i), t(v_j)\,). \qquad (8.5)\end{aligned}$$

Modularity

This measure is called *modularity* [211]. The maximum modularity value for a network depends on the number of nodes of the same type and degree. The maximum occurs when all edges are connecting nodes of the same type (i.e., when $A_{ij} = 1$, $\delta(\, t(v_i), t(v_j)\,) = 1$). We can normalize modularity by dividing it by the maximum it can take:

$$\begin{aligned} Q_{\text{normalized}} &= \frac{Q}{Q_{\max}} \qquad (8.6)\\ &= \frac{\frac{1}{2m}\sum_{ij}(\, A_{ij} - \frac{d_i d_j}{2m}\,)\,\delta(\, t(v_i), t(v_j)\,)}{\max[\frac{1}{2m}\sum_{ij} A_{ij}\delta(\, t(v_i), t(v_j)\,) - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\delta(\, t(v_i), t(v_j)\,)]} \qquad (8.7)\\ &= \frac{\frac{1}{2m}\sum_{ij}(\, A_{ij} - \frac{d_i d_j}{2m}\,)\,\delta(\, t(v_i), t(v_j)\,)}{\frac{1}{2m}2m - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}\,\delta(\, t(v_i), t(v_j)\,)} \qquad (8.8)\\ &= \frac{\sum_{ij}(\, A_{ij} - \frac{d_i d_j}{2m}\,)\,\delta(\, t(v_i), t(v_j)\,)}{2m - \sum_{ij}\frac{d_i d_j}{2m}\,\delta(\, t(v_i), t(v_j)\,)}. \qquad (8.9)\end{aligned}$$

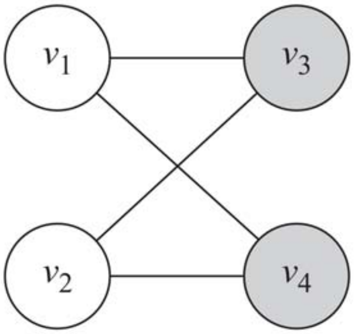


Figure 8.3: A Modularity Example for a Bipartite Graph.

Modularity can be simplified using a matrix format. Let $\Delta \in \mathbb{R}^{n\times k}$ denote the indicator matrix and let k denote the number of types,

$$\Delta_{x,k} = \begin{cases} 1, & \text{if } t(x) = k; \\ 0, & \text{if } t(x) \neq k \end{cases} \tag{8.10}$$

Note that δ function can be reformulated using the indicator matrix:

$$\delta(\ t(v_i), t(v_j)\ ) = \sum_k \Delta_{v_i,k}\Delta_{v_j,k}. \tag{8.11}$$

Therefore, $(\Delta\Delta^T)_{i,j} = \delta(t(v_i), t(v_j))$. Let $B = A - \mathbf{d}\mathbf{d}^T/2m$ denote the modularity matrix where $\mathbf{d} \in \mathbb{R}^{n\times 1}$ is the degree vector for all nodes. Given that the trace of multiplication of two matrices X and Y^T is $Tr(XY^T) = \sum_{i,j} X_{i,j}Y_{i,j}$ and $Tr(XY) = Tr(YX)$, modularity can be reformulated as

$$\begin{aligned} Q &= \frac{1}{2m}\sum_{ij} \underbrace{(\ A_{ij} - \frac{d_i d_j}{2m}\ )}_{B_{ij}}\ \underbrace{\delta(\ t(v_i), t(v_j)\ )}_{(\Delta\Delta^T)_{i,j}} = \frac{1}{2m}\mathrm{Tr}(B\Delta\Delta^T) \\ &= \frac{1}{2m}\mathrm{Tr}(\Delta^T B\Delta). \end{aligned} \tag{8.12}$$

Example 8.1. *Consider the bipartite graph in Figure 8.3. For this bipartite graph,*

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \end{bmatrix}, \quad \Delta = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{d} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 2 \end{bmatrix}, m = 4. \tag{8.13}$$

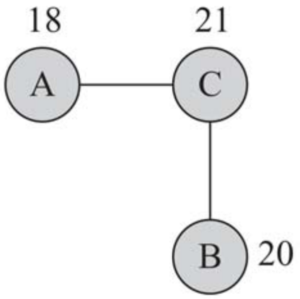


Figure 8.4: A Correlation Example.

Therefore, matrix B is

$$B = A - \mathbf{dd}^T/2m = \begin{bmatrix} -0.5 & -0.5 & 0.5 & 0.5 \\ -0.5 & -0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \\ 0.5 & 0.5 & -0.5 & -0.5 \end{bmatrix}. \tag{8.14}$$

The modularity value Q is

$$\frac{1}{2m} Tr(\Delta^T B \Delta) = -0.5. \tag{8.15}$$

In this example, all edges are between nodes of different color. In other words, the number of edges between nodes of the ***same color*** *is less than the expected number of edges between them. Therefore, the modularity value is negative.*

8.1.2 Measuring Assortativity for Ordinal Attributes

A common measure for analyzing the relationship between two variables with ordinal values is covariance. Covariance describes how two variables change with respect to each other. In our case, we are interested in how correlated, the attribute values of nodes connected via edges are. Let x_i be the ordinal attribute value associated with node v_i. In Figure 8.4, for node c, the value associated is $x_c = 21$.

Covariance

We construct two variables X_L and X_R, where for any edge (v_i, v_j) we assume that x_i is observed from variable X_L and x_j is observed from variable

X_R. For Figure 8.4,

$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix}, \quad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix}. \tag{8.16}$$

In other words, X_L represents the ordinal values associated with the left node of the edges, and X_R represents the values associated with the right node of the edges. Our problem is therefore reduced to computing the covariance between variables X_L and X_R. Note that since we are considering an undirected graph, both edges (v_i, v_j) and (v_j, v_i) exist; therefore, x_i and x_j are observed in both X_L and X_R. Thus, X_L and X_R include the same set of values but in a different order. This implies that X_L and X_R have the same mean and standard deviation.

$$\mathbf{E}(X_L) = \mathbf{E}(X_R), \tag{8.17}$$

$$\sigma(X_L) = \sigma(X_R). \tag{8.18}$$

Since we have m edges and each edge appears twice for the undirected graph, then X_L and X_R have $2m$ elements. Each value x_i appears d_i times since it appears as endpoints of d_i edges. The covariance between X_L and X_R is

$$\begin{aligned} \sigma(X_L, X_R) &= \mathbf{E}[(X_L - \mathbf{E}[X_L])(X_R - \mathbf{E}[X_R])] \\ &= \mathbf{E}[X_L X_R - X_L \mathbf{E}[X_R] - \mathbf{E}[X_L] X_R + \mathbf{E}[X_L]\mathbf{E}[X_R]] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R] + \mathbf{E}[X_L]\mathbf{E}[X_R] \\ &= \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R]. \end{aligned} \tag{8.19}$$

$\mathbf{E}(X_L)$ is the mean (expected value) of variable X_L, and $\mathbf{E}(X_L X_R)$ is the mean of the multiplication of X_L and X_R. In our setting and following Equation 8.17, these expectations are as follows:

$$E(X_L) = E(X_R) = \frac{\sum_i (X_L)_i}{2m} = \frac{\sum_i d_i x_i}{2m} \tag{8.20}$$

$$E(X_L X_R) = \frac{1}{2m} \sum_i (X_L)_i (X_R)_i = \frac{\sum_{ij} A_{ij} x_i x_j}{2m}. \tag{8.21}$$

By plugging Equations 8.20 and 8.21 into Equation 8.19, the covariance between X_L and X_R is

$$\sigma(X_L, X_R) = \mathbf{E}[X_L X_R] - \mathbf{E}[X_L]\mathbf{E}[X_R]$$

$$
\begin{aligned}
&= \frac{\sum_{ij} A_{ij}x_i x_j}{2m} - \frac{\sum_{ij} d_i d_j x_i x_j}{(2m)^2} \\
&= \frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})x_i x_j. \qquad (8.22)
\end{aligned}
$$

Similar to modularity (Section 8.1.1), we can normalize covariance. Pearson correlation $\rho(X_L, X_R)$ is the normalized version of covariance:

Pearson Correlation

$$\rho(X_L, X_R) = \frac{\sigma(X_L, X_R)}{\sigma(X_L)\sigma(X_R)}. \qquad (8.23)$$

From Equation 8.18, $\sigma(X_L) = \sigma(X_R)$; thus,

$$
\begin{aligned}
\rho(X_L, X_R) &= \frac{\sigma(X_L, X_R)}{\sigma(X_L)^2}, \\
&= \frac{\frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})x_i x_j}{\mathbf{E}[(X_L)^2] - (\mathbf{E}[X_L])^2} \\
&= \frac{\frac{1}{2m}\sum_{ij}(A_{ij} - \frac{d_i d_j}{2m})x_i x_j}{\frac{1}{2m}\sum_{ij} A_{ij}x_i^2 - \frac{1}{2m}\sum_{ij}\frac{d_i d_j}{2m}x_i x_j}. \qquad (8.24)
\end{aligned}
$$

Note the similarity between Equations 8.9 and 8.24. Although modularity is used for nominal attributes and correlation for ordinal attributes, the major difference between the two equations is that the δ function in modularity is replaced by $x_i x_j$ in the correlation equation.

Example 8.2. *Consider Figure 8.4 with values demonstrating the attributes associated with each node. Since this graph is undirected, we have the following edges:*

$$E = \{(a, c), (c, a), (c, b), (b, c)\}. \qquad (8.25)$$

The correlation is between the values associated with the endpoints of the edges. Consider X_L as the value of the left end of an edge and X_R as the value of the right end of an edge:

$$X_L = \begin{bmatrix} 18 \\ 21 \\ 21 \\ 20 \end{bmatrix}, \quad X_R = \begin{bmatrix} 21 \\ 18 \\ 20 \\ 21 \end{bmatrix} \qquad (8.26)$$

The correlation between these two variables is $\rho(X_L, X_R) = -0.67$.

8.2 Influence

Influence[3] is "the act or power of producing an effect without apparent exertion of force or direct exercise of command." In this section, we discuss influence and, in particular, how we can (1) measure influence in social media and (2) design models that detail how individuals influence one another in social media.

8.2.1 Measuring Influence

Prediction-Based Influence Measures

Influence can be measured based on (1) *prediction* or (2) *observation*.

Prediction-Based Measures. In prediction-based measurement, we assume that an individual's attribute or the way she is situated in the network predicts how influential she *will be*. For instance, we can assume that the gregariousness (e.g., number of friends) of an individual is correlated with how influential she *will be*. Therefore, it is natural to use any of the centrality measures discussed in Chapter 3 for prediction-based influence measurements. Examples of such centrality measures include PageRank and degree centrality. In fact, many of these centrality measures were introduced as influence-measuring techniques. For instance, on Twitter, in-degree (number of followers) is a common attribute for measuring influence. Since these methods were covered in-depth in that chapter, in this section we focus on observational techniques.

Observation-based Influence Measures

Observation-Based Measures. In observation-based measures, we quantify the influence of an individual by measuring the amount of influence attributed to him. An individual can influence differently in diverse settings, and so, depending on the context, the observation-based measuring of influence changes. We next describe three different settings and how influence can be measured in each.

1. **When an individual is the role model**. This happens in the case of individuals in the fashion industry, teachers, and celebrities. In this case, **the size of the audience that has been influenced** due to that fashion, charisma, or the like could act as an accurate measure. A

[3] As defined by the Merriam-Webster dictionary.

local grade-school teacher has a tremendous influence over a class of students, whereas Gandhi influenced millions.

2. **When an individual spreads information**. This scenario is more likely when a piece of information, an epidemic, or a product is being spread in a network. In this case, **the size of the cascade** – that is, the number of hops the information traveled – or **the population affected**, or the **rate at which population gets influenced** is considered a measure.

3. **When an individual's participation increases the value of an item or action.** As in the case of diffusion of innovations (see Chapter 7), often when individuals perform actions such as buying a product, they increase the value of the product for other individuals. For example, the first individual who bought a fax machine had no one to send faxes to. The second individual who bought a fax machine increased its value for the first individual. So, the **increase (or rate of increase) in the value of an item or action** (such as buying a product) is often used as a measure.

Case Studies for Measuring Influence in Social Media

This section provides examples of measuring influence in the blogosphere and on the microblogging site Twitter. These techniques can be adapted to other social media sites, as well.

Measuring Social Influence in the Blogosphere

The goal of measuring influence in the blogosphere is to identify influential bloggers. Due to the limited time that individuals have, following the influentials is often necessary for fast access to interesting news. One common measure for quantifying the influence of bloggers is to use in-degree centrality: the number of (in-)links that point to the blog. However, because of the sparsity of in-links, more detailed analysis is required to measure influence in the blogosphere.

In their book, *The Influentials: One American in Ten Tells the Other Nine How to Vote, Where to Eat, and What to Buy*. Keller and Berry [145] argue that the influentials are individuals who (1) are recognized by others, (2) whose activities result in follow-up activities, (3) have novel perspectives, and (4) are eloquent.

To address these issues, Agarwal et al. [6] proposed the *iFinder* system to measure influence of blogposts and to identify influential bloggers. In particular, for each one of these four characteristics and a blogpost p, they approximate the characteristic by collecting specific blogpost's attributes:

1. **Recognition**. Recognition for a blogpost can be approximated by the links that point to the blogpost (in-links). Let $\mathcal{I}_p$ denote the set of in-links that point to blogpost p.

2. **Activity Generation**. Activity generated by a blogpost can be estimated using the number of comments that p receives. Let c_p denote the number of comments that blogpost p receives.

3. **Novelty.** The blogpost's novelty is inversely correlated with the number of references a blogpost employs. In particular the more citations a blogpost has, the less novel it is. Let $\mathcal{O}_p$ denote the set of out-links for blogpost p.

4. **Eloquence.** Eloquence can be estimated by the length of the blogpost. Given the informal nature of blogs and the bloggers' tendency to write short blogposts, longer blogposts are commonly believed to be more eloquent. So, the length of a blogpost l_p can be employed as a measure of eloquence.

Given these approximations for each one of these characteristics, we can design a measure of influence for each blogpost. Since the number of out-links inversely affects the influence of a blogpost and the number of in-links increases it, we construct an influence graph, or *i-graph*, where blogposts are nodes and influence flows through the nodes. The amount of this *influence flow* for each post p can be characterized as

Influence Flow

$$InfluenceFlow(p) = w_{\text{in}} \sum_{m=1}^{|\mathcal{I}_p|} I(P_m) - w_{\text{out}} \sum_{n=1}^{|\mathcal{O}_p|} I(P_n), \tag{8.27}$$

where $I(.)$ denotes the influence of a blogpost and w_{in} and w_{out} are the weights that adjust the contribution of in- and out-links, respectively. In this equation, P_m's are blogposts that point to post p, and P_n's are blogposts that are referred to in post p. Influence flow describes a measure that only accounts for in-links (recognition) and out-links (novelty). To account for the other two factors, we design the influence of a blogpost p as

$$I(p) = w_{\text{length}} l_p (w_{\text{comment}} c_p + \mathit{InfluenceFlow}(p)). \tag{8.28}$$

Here, w_{length} is the weight for the length of blogpost[4]. w_{comment} describes how the number of comments is weighted. Note that the four weights w_{in}, w_{out}, w_{comments}, and w_{length} need to be tuned to make the model more accurate. This tuning can be done by a variety of techniques. For instance, we can use a test system where the influential posts are already known (labeled data) to tune them.[5] Finally, a blogger's influence index (*iIndex*) can be defined as the maximum influence value among all his or her N blogposts,

$$\mathit{iIndex} = \max_{p_n \in N} I(p_n). \tag{8.29}$$

Computing *iIndex* for a set of bloggers over all their blogposts can help identify and rank influential bloggers in a system.

Measuring Social Influence on Twitter. On Twitter, a microblogging platform, users receive tweets from other users by *following* them. Intuitively, we can think of the number of followers as a measure of influence (in-degree centrality). In particular, three measures are frequently used to quantify influence in Twitter,

1. **In-degree**: the number of users following a person on Twitter. As discussed, the number of individuals who are interested in someone's tweets (i.e., followers) is commonly used as an influence measure on Twitter. In-degree denotes the "audience size" of an individual.

2. **Number of mentions**: the number of times an individual is mentioned in tweets. Mentioning an individual with a `username` handle is performed by including `@username` in a tweet. The number of times an individual is mentioned can be used as an influence measure. The number of mentions denotes the "ability in engaging others in conversation" [50].

[4] In the original paper, the authors utilize a weight function instead. Here, for clarity, we use coefficients for all parameters.

[5] Note that Equation 8.28 is defined recursively, because $I(p)$ depends on *InfluenceFlow* and that, in turn, depends on $I(p)$ (Equation 8.27). Therefore, to estimate $I(p)$, we can use iterative methods where we start with an initial value for $I(p)$ and compute until convergence.

Table 8.1: Rank Correlation between Top 10% of Influentials for Different Measures on Twitter

Measures	Correlation Value
In-degree vs. retweets	0.122
In-degree vs. mentions	0.286
Retweets vs. mentions	0.638

3. **Number of retweets**: the number of times tweets of a user are retweeted. Individuals on Twitter have the opportunity to forward tweets to a broader audience via the retweet capability. Clearly, the more one's tweets are retweeted, the more likely one is influential. The number of retweets indicates an individual's ability to generate content that is worth being passed along.

Each one of these measures by itself can be used to identify influential users in Twitter. This can be done by utilizing the measure for each individual and then ranking individuals based on their measured influence value. Contrary to public perception, the number of followers is considered an inaccurate measure compared to the other two. This is shown in [50], where the authors ranked individuals on Twitter independently based on these three measures. To see if they are correlated or redundant, they compared ranks of individuals across three measures using rank correlation measures. One such measure is the Spearman's rank correlation coefficient,

Spearman's Rank Correlation

$$\rho = 1 - \frac{6\sum_{i=1}^{n}(m_1^i - m_2^i)^2}{n^3 - n}, \tag{8.30}$$

where m_1^i and m_2^i are ranks of individual i based on measures m_1 and m_2, and n is the total number of users. Spearman's rank correlation is the Pearson correlation coefficient for ordinal variables that represent ranks (i.e., takes values between 1. . . n); hence, the value is in range [−1,1]. Their findings suggest that popular users (users with high in-degree) do not necessarily have high ranks in terms of number of retweets or mentions. This can be observed in Table 8.1, which shows the Spearman's correlation between the top 10% influentials for each measure.

8.2.2 Modeling Influence

In influence modeling, the goal is to design models that can explain how individuals influence one another. Given the nature of social media, it is safe to assume that influence takes place among connected individuals. At times, this network is observable (explicit networks), and at others times, it is unobservable (implicit networks). For instance, in referral networks, where people refer others to join an online service on social media, the network of referrals is often observable. In contrast, people are influenced to buy products, and in most cases, the seller has no information on who referred the buyer, but does have approximate estimates on the number of products sold over time. In the observable (explicit) network, we resort to threshold models such as the linear threshold model (LTM) to model influence; in implicit networks, we can employ methods such as the linear influence model (LIM) that take the number of individuals who get influenced at different times as input (e.g., the number of buyers per week).

Linear Threshold Model (LTM)

Modeling Influence in Explicit Networks

Threshold models are simple yet effective methods for modeling influence in explicit networks. In these models, nodes make decision based on the number or the fraction (the threshold) of their neighbors (or incoming neighbors in a directed graph) who have already decided to make the same decision. Threshold models were employed in the literature as early as the 1970s in the works of Granovetter [109] and Schelling [251]. Using a threshold model, Schelling demonstrated that minor local preferences in having neighbors of the same color leads to global racial segregation.

A *linear threshold model (LTM)* is an example of a threshold model. Assume a weighted directed graph where nodes v_j and v_i are connected with weight $w_{j,i} \geq 0$. This weight denotes how much node v_j can affect node v_i's decision. We also assume

$$\sum_{v_j \in N_{\text{in}}(v_i)} w_{j,i} \leq 1, \tag{8.31}$$

where $N_{\text{in}}(v_i)$ denotes the incoming neighbors of node v_i. In a linear threshold model, each node v_i is assigned a threshold θ_i such that when the amount of influence exerted toward v_i by its active incoming neighbors is

Algorithm 8.1 Linear Threshold Model (LTM)

```
Require: Graph $G(V, E)$, set of initial activated nodes $A_0$
return Final set of activated nodes $A_\infty$
i=0;
Uniformly assign random thresholds $\theta_v$ from the interval [0, 1];
while $i = 0$ or $(A_{i-1} \neq A_i, i \geq 1)$ do
    $A_{i+1} = A_i$
    inactive = $V - A_i$;
    for all $v \in$ inactive do
        if $\sum_{j \text{ connected to } v, j \in A_i} w_{j,v} \geq \theta_v$. then
            activate $v$;
            $A_{i+1} = A_{i+1} \cup \{v\}$;
        end if
    end for
    $i = i + 1$;
end while
$A_\infty = A_i$;
Return $A_\infty$;
```

more than θ_i, then v_i becomes active, if still inactive. Thus, for v_i to become active at time t, we should have

$$\sum_{v_j \in N_{\text{in}}(v_i), v_j \in A_{t-1}} w_{j,i} \geq \theta_i, \tag{8.32}$$

where A_{t-1} denotes the set of active nodes at the end of time $t - 1$. The threshold values are generally assigned uniformly at random to nodes from the interval [0,1]. Note that the threshold θ_i defines how resistant to change node v_i is: a very small θ_i value might indicate that a small change in the activity of v_i's neighborhood results in v_i becoming active and a large θ_i shows that v_i resists changes.

Provided a set of initial active nodes A_0 and a graph, the LTM algorithm is shown in Algorithm 8.1. In each step, for all inactive nodes, the condition in Equation 8.32 is checked, and if it is satisfied, the node becomes active. The process ends when no more nodes can be activated. Once θ thresholds are fixed, the process is deterministic and will always converge to the same state.

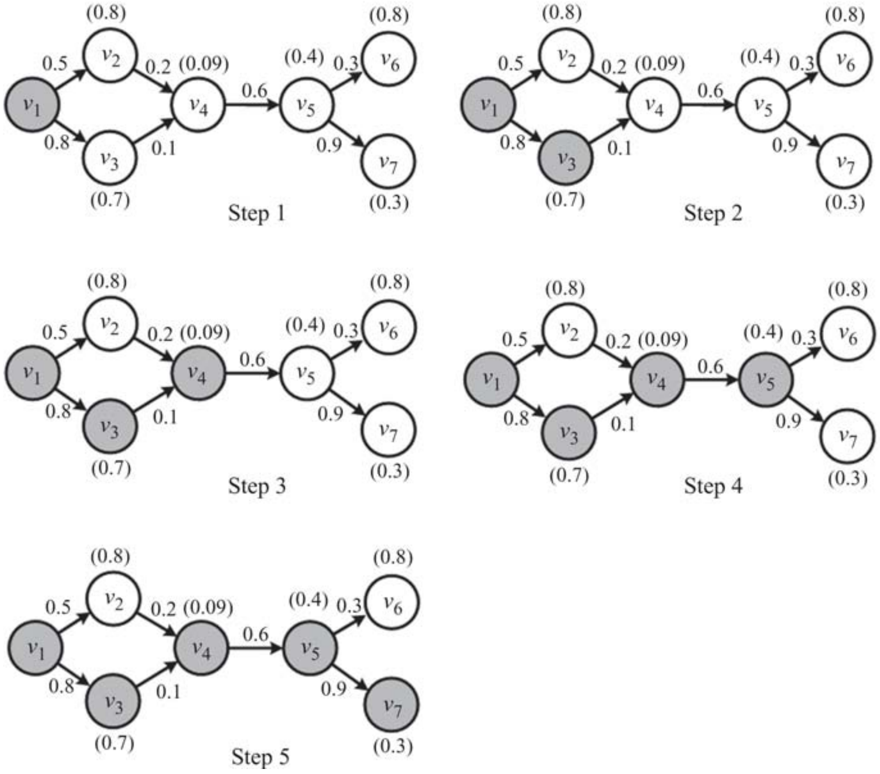


Figure 8.5: Linear Threshold Model (LTM) Simulation. The values attached to nodes denote thresholds θ_i, and the values on the edges represent weights $w_{i,j}$.

Example 8.3. *Consider the graph in Figure 8.5. Values attached to nodes represent the LTM thresholds, and edge values represent the weights. At time 0, node v_1 is activated. At time 2, both nodes v_2 and v_3 receive influence from node v_1. Node v_2 is not activated since 0.5 < 0.8 and node v_3 is activated since 0.8 > 0.7. Similarly, the process continues and then stops with five activated nodes.*

Modeling Influence in Implicit Networks

An implicit network is one where the influence spreads over edges in the network; however, unlike the explicit model, we cannot observe the individuals (the influentials) who are responsible for influencing others, but only those who get influenced. In other words, the information available

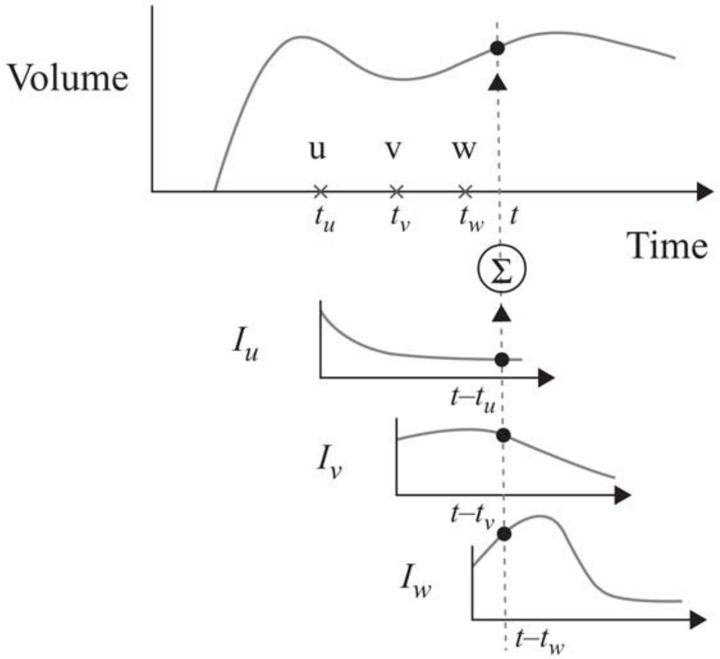


Figure 8.6: The Size of the Influenced Population as a Summation of Individuals Influenced by Activated Individuals (from [306]).

is the set of influenced population $P(t)$ at any time and the time t_u, when each individual u gets initially influenced (activated). We assume that any influenced individual u can influence $I(u, t)$ number of non-influenced (inactive) individuals after t time steps. We call $I(., .)$ the influence function. Assuming discrete time steps, we can formulate the size of the influenced population $|P(t)|$:

$$|P(t)| = \sum_{u \in P(t)} I(u, t - t_u). \tag{8.33}$$

Figure 8.6 shows how the model performs. Individuals u, v, and w are activated at time steps t_u, t_v, and t_w, respectively. At time t, the total number of influenced individuals is a summation of influence functions I_u, I_v, and I_w at time steps $t - t_u$, $t - t_v$, and $t - t_w$, respectively. Our goal is to estimate $I(., .)$ given activation times and the number of influenced individuals at all times. A simple approach is to utilize a probability distribution to estimate I function. For instance, we can employ the power-law distribution to estimate influence. In this case, $I(u,t) = c_u(t - t_u)^{-\alpha_u}$, where we estimate coefficients c_u and α_u for any u by methods such as *maximum likelihood* estimation (see [205] for more details).

This is called the *parametric* estimation, and the method assumes that all users influence others in the same parametric form. A more flexible approach is to assume a nonparametric function and estimate the influence function's form. This approach was first introduced as the linear influence model (LIM) [306].

Linear Influence Model (LIM)

In LIM, we extend our formulation by assuming that nodes get deactivated over time and then no longer influence others. Let $A(u,t) = 1$ denote that node u is active at time t, and $A(u,t) = 0$ denote that node u is either deactivated or still not influenced. Following a network notation and assuming that $|V|$ is the total size of the population and T is the last time step, we can reformulate Equation 8.33 for $|P(t)|$ as

$$|P(t)| = \sum_{u=1}^{|V|} \sum_{t=1}^{T} A(u,t)I(u,t), \tag{8.34}$$

or equivalently in matrix form,

$$P = AI. \tag{8.35}$$

It is common to assume that individuals can only activate other individuals and cannot stop others from becoming activated. Hence, negative values for influence do not make sense; therefore, we would like measured influence values to be positive $I \geq 0$,

$$\begin{aligned} \text{minimize} \quad & ||P - AI||_2^2 && (8.36) \\ \text{subject to} \quad & I \geq 0. && (8.37) \end{aligned}$$

This formulation is similar to regression coefficients computation outlined in Chapter 5, where we compute a least square estimate of I; however, this formulation cannot be solved using regression techniques studied earlier because, in regression, computed I values can become negative. In practice, this formulation can be solved using non-negative least square methods (see [164] for details).

8.3 Homophily

Homophily is the tendency of similar individuals to become friends. It happens on a daily basis in social media and is clearly observable in social networking sites where befriending can explicitly take place. The well-known saying, "birds of a feather flock together," is frequently quoted when discussing homophily. Unlike influence, where an influential influences others, in homophily, **two** similar individuals decide to get connected.

8.3.1 Measuring Homophily

Homophily is the linking of two individuals due to their similarity and leads to assortative networks over time. To measure homophily, we measure how the assortativity of the network has changed over time.[6] Consider two snapshots of a network $G_{t_1}(V, E_{t_1})$ and $G_{t_2}(V, E_{t_2})$ at times t_1 and t_2, respectively, where $t_2 > t_1$. Without loss of generality, we assume that the number of nodes is fixed and only edges connecting these nodes change (i.e., are added or removed).

When dealing with nominal attributes, the homophily index is defined as

$$H = Q^{t_2}_{\text{normalized}} - Q^{t_1}_{\text{normalized}}, \tag{8.38}$$

where $Q_{\text{normalized}}$ is defined in Equation 8.9. Similarly, for ordinal attributes, the homophily index can be defined as the change in the Pearson correlation (Equation 8.24):

$$H = \rho^{t_2} - \rho^{t_1}. \tag{8.39}$$

8.3.2 Modeling Homophily

Homophily can be modeled using a variation of the independent cascade model discussed in Chapter 7. In this variation, at each time step a single node gets activated, and the activated node gets a chance of getting connected to other nodes due to homophily. In other words, if the activated node finds other nodes in the network similar enough (i.e., their similarity

[6] Note that we have assumed that homophily is the leading social force in the network and that it leads to its assortativity change. This assumption is often strong for social networks because other social forces act in these networks.

Algorithm 8.2 Homophily Model

```
Require: Graph $G(V,E)$, $E = \emptyset$, similarities $sim(v,u)$
return  Set of edges $E$
for all $v \in V$ do
    $\theta_v$ = generate a random number in [0,1];
    for all $(v,u) \notin E$ do
        if $\theta_v < sim(v,u)$ then
            $E = E \cup (v,u)$;
        end if
    end for
end for
Return $E$;
```

is higher than some tolerance value), it connects to them via an edge. A node once activated has no chance of getting activated again.

Modeling homophily is outlined in Algorithm 8.2. Let $sim(u,v)$ denote the similarity between nodes u and v. When a node gets activated, we generate a random tolerance value for the node v between 0 and 1. Alternatively, we can set this tolerance to some predefined value. The tolerance value defines the minimum similarity that node v tolerates for connecting to other nodes. Then, for any likely edge (v,u) that is still not in the edge set, if the similarity is more than the tolerance: $sim(v,u) > \theta_v$, the edge (v,u) is added. The process continues until all nodes are activated.

The model can be used in two different scenarios. First, given a network in which assortativity is attributed to homophily, we can estimate tolerance values for all nodes. To estimate tolerance values, we can simulate the homophily model in Algorithm 8.2 on the given network (by removing all its edges) with different tolerance values. We can then compare the assortativity of the simulated network and the given network. By finding the simulated network that best fits the given network (i.e., has the closest assortativity value to the given network's assortativity), we can determine the tolerance values for individuals. Second, when a network is given and the source of assortativity is unknown, we can estimate how much of the observed assortativity can be attributed to homophily. To measure assortativity due to homophily, we can simulate homophily on the given network by removing edges. The distance between the assortativity measured on the simulated network and the given network explains how much of the observed assortativity is due to homophily. The smaller this distance, the higher the effect of homophily in generating the observed assortativity.

8.4 Distinguishing Influence and Homophily

We are often interested in understanding which social force (influence or homophily) resulted in an assortative network. To distinguish between an influence-based assortativity or homophily-based one, statistical tests can be used. In this section, we discuss three tests: the shuffle test, the edge-reversal test, and the randomization test. The first two can detect whether influence exists in a network or not, but are incapable of detecting homophily. The last one, however, can distinguish influence and homophily. Note that in all these tests, we assume that several temporal snapshots of the dataset are available (like the LIM model) where we know exactly when each node is activated, when edges are formed, or when attributes are changed.

8.4.1 Shuffle Test

The shuffle test was originally introduced by Anagnostopoulos et al. [10]. The basic idea behind the shuffle test comes from the fact that influence is temporal. In other words, when u influences v, then v should have been activated after u. So, in the shuffle test, we define a temporal assortativity measure. We assume that if there is no influence, then a shuffling of the activation time stamps should not affect the temporal assortativity measurement.

Social Correlation

In this temporal assortativity measure, called *social correlation*, the probability of activating a node v depends on a, the number of already active friends it has. This activation probability is calculated using a logistic function,[7]

$$p(a) = \frac{e^{\alpha a+\beta}}{1+e^{\alpha a+\beta}}, \tag{8.40}$$

[7] In the original paper, instead of a, the authors use $\ln(a+1)$ as the variable. This helps remove the effect of a power-law distribution in the number of activated friends. Here, for simplicity, we use the nonlogarithmic form.

or equivalently,

$$\ln(\frac{p(a)}{1 - p(a)}) = \alpha a + \beta, \tag{8.41}$$

where α measures the social correlation and β denotes the activation bias. For computing the number of already active nodes of an individual, we need to know the activation time stamps of the nodes.

Let $y_{a,t}$ denote the number of individuals who became activated at time t and had a active friends and let $n_{a,t}$ denote the ones who had a active friends but did not get activated at time t. Let $y_a = \sum_t y_{a,t}$ and $n_a = \sum_t n_{a,t}$. We define the likelihood function as

$$\prod_a p(a)^{y_a}(1 - p(a))^{n_a}. \tag{8.42}$$

To estimate α and β, we find their values such that the likelihood function denoted in Equation 8.42 is maximized. Unfortunately, there is no closed-form solution, but there exist software packages that can efficiently compute the solution to this optimization.[8]

Let t_u denote the activation time (when a node is first influenced) of node u. When activated node u influences nonactivated node v, and v is activated, then we have $t_u < t_v$. Hence, when temporal information is available about who activated whom, we see that influenced nodes are activated at a later time than those who influenced them. Now, if there is no influence in the network, we can randomly shuffle the activation time stamps, and the predicted α should not change drastically. So, if we shuffle activation time stamps and compute the correlation coefficient α' and its value is close to the α computed in the original unshuffled dataset (i.e., $|\alpha - \alpha'|$ is small), then the network does not exhibit signs of social influence.

8.4.2 Edge-Reversal Test

The edge-reversal test introduced by Christakis and Fowler [54] follows a similar approach as the shuffle test. If influence resulted in activation, then the direction of edges should be important (who influenced whom). So, we can reverse the direction of edges, and if there is no social influence in the network, then the value of social correlation α, as defined in Section 8.4.1, should not change dramatically.

[8] Note that maximizing this term is equivalent to maximizing the logarithm; this is where Equation 8.41 comes into play.

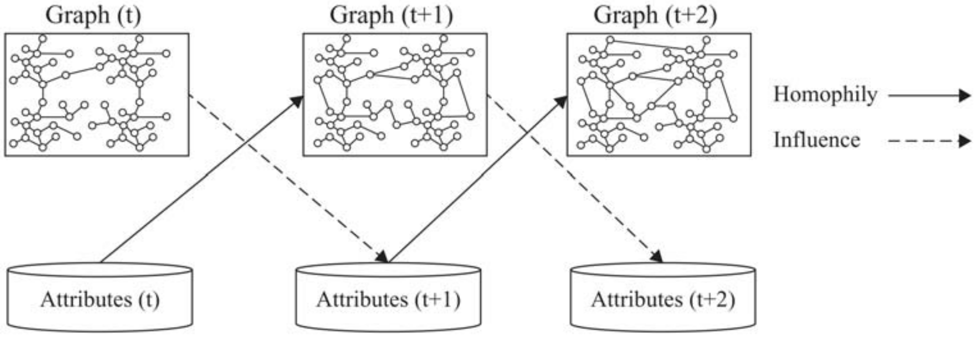


Figure 8.7: The Effect of Influence and Homophily on Attributes and Links over Time (reproduced from [161]).

8.4.3 Randomization Test

Unlike the other two tests, the randomization test [161] is capable of detecting both influence and homophily in networks. Let X denote the attributes associated with nodes (age, gender, location, etc.) and X_t denote the attributes at time t. Let X^i denote attributes of node v_i. As mentioned before, in influence, individuals already linked to one another change their attributes (e.g., a user changes habits), whereas in homophily, attributes do not change but connections are formed due to similarity. Figure 8.7 demonstrates the effect of influence and homophily in a network over time.

The assumption is that, if influence or homophily happens in a network, then networks become more assortative. Let $A(G_t, X_t)$ denote the assortativity of network G and attributes X at time t. Then, the network becomes more assortative at time $t + 1$ if

$$A(G_{t+1}, X_{t+1}) - A(G_t, X_t) > 0. \tag{8.43}$$

Influence Gain and Homophily Gain

Now, we can assume that part of this assortativity is due to influence if the *influence gain* $G_{\text{Influence}}$ is positive,

$$G_{\text{Influence}}(t) = A(G_t, X_{t+1}) - A(G_t, X_t) > 0, \tag{8.44}$$

and part is due to homophily if we have positive *homophily gain* $G_{\text{Homophily}}$:

$$G_{\text{Homophily}}(t) = A(G_{t+1}, X_t) - A(G_t, X_t) > 0. \tag{8.45}$$

Algorithm 8.3 Influence Significance Test

```
Require: $G_t$, $G_{t+1}$, $X_t$, $X_{t+1}$, number of randomized runs $n$, $\alpha$
return Significance
$g_0 = G_{Influence}(t)$;
for all $1 \leq i \leq n$ do
    $XR^i_{t+1} = randomize_I(X_t, X_{t+1})$;
    $g_i = A(G_t, XR^i_{t+1}) - A(G_t, X_t)$;
end for
if $g_0$ larger than $(1 - \alpha/2)\%$ of values in $\{g_i\}_{i=1}^n$ then
    return significant;
else if $g_0$ smaller than $\alpha/2\%$ of values in $\{g_i\}_{i=1}^n$ then
    return significant;
else
    return insignificant;
end if
```

Note that X_{t+1} denotes the changes in attributes, and G_{t+1} denotes the changes in links in the network (new friendships formed). In randomization tests, one determines whether changes in $A(G_t, X_{t+1}) - A(G_t, X_t)$ (influence), or $A(G_{t+1}, X_t) - A(G_t, X_t)$ (homophily), are significant or not. To detect change significance, we use the influence significance test and homophily significance test algorithms outlined in Algorithms 8.3 and 8.4, respectively. The influence significance algorithm starts with computing influence gain, which is the assortativity difference observed due to influence (g_0). It then forms a random attribute set at time $t+1$ (null-hypotheses), assuming that attributes changed randomly at $t + 1$ and not due to influence. This random attribute set XR^i_{t+1} is formed from X_{t+1} by making sure that effects of influence in changing attributes are removed.

Influence Significance Test

For instance, assume two users u and v are connected at time t, and u has hobby `movies` at time t and v does not have this hobby listed at time t. Now, assuming there is an influence of u over v, so that at time $t + 1$, v adds `movies` to her set of hobbies. In other words, `movies` $\notin X^v_t$ and `movies` $\in X^v_{t+1}$. To remove this influence, we can construct XR^i_{t+1} by removing `movies` from the hobbies of v at time $t + 1$ and adding some random hobby such as `reading`, which is $\notin X^u_t$ and $\notin X^v_t$, to the list of hobbies of v at time $t + 1$ in XR^i_{t+1}. This guarantees that the randomized XR^i_{t+1} constructed has no sign of influence. We construct this randomized set n times; this set is

Algorithm 8.4 Homophily Significance Test

```
Require: $G_t$, $G_{t+1}$, $X_t$, $X_{t+1}$, number of randomized runs $n$, $\alpha$
return Significance
$g_0 = G_{Homophily}(t)$;
for all $1 \leq i \leq n$ do
    $GR^i_{t+1} = randomize_H(G_t, G_{t+1})$;
    $g_i = A(GR^i_{t+1}, X_t) - A(G_t, X_t)$;
end for
if $g_0$ larger than $(1 - \alpha/2)\%$ of values in $\{g_i\}^n_{i=1}$ then
    return significant;
else if $g_0$ smaller than $\alpha/2\%$ of values in $\{g_i\}^n_{i=1}$ then
    return significant;
else
    return insignificant;
end if
```

then used to compute influence gains $\{g_i\}^n_{i=1}$. Obviously, the more distant g_0 is from these gains, the more significant influence is. We can assume that whenever g_0 is smaller than $\alpha/2\%$ (or larger than $1 - \alpha/2\%$) of $\{g_i\}^n_{i=1}$ values, it is significant. The value of α is set empirically.

Homophily Significance Test

Similarly, in the homophily significance test, we compute the original homophily gain and construct random graph links GR^i_{t+1} at time $t+1$, such that no homophily effect is exhibited in how links are formed. To perform this for any two (randomly selected) links e_{ij} and e_{kl} formed in the original G_{t+1} graph, we form edges e_{il} and e_{kj} in GR^i_{t+1}. This is to make sure that the homophily effect is removed and that the degrees in GR^i_{t+1} are equal to that of G_{t+1}.

8.5 Summary

Individuals are driven by different social forces across social media. Two such important forces are influence and homophily.

In influence, an individual's actions induce her friends to act in a similar fashion. In other words, influence makes friends more similar. Homophily is the tendency for similar individuals to befriend each other. Both influence and homophily result in networks where similar individuals are connected to each other. These are assortative networks. To estimate the assortativity of networks, we use different measures depending on the attribute type that is tested for similarity. We discussed modularity for nominal attributes and correlation for ordinal ones.

Influence can be quantified via different measures. Some are prediction-based, where the measure assumes that some attributes can accurately predict how influential an individual will be, such as with in-degree. Others are observation-based, where the influence score is assigned to an individual based on some history, such as how many individuals he or she has influenced. We also presented case studies for measuring influence in the blogosphere and on Twitter.

Influence is modeled differently depending on the visibility of the network. When network information is available, we employ threshold models such as the linear threshold model (LTM), and when network information is not available, we estimate influence rates using the linear influence model (LIM). Similarly, homophily can be measured by computing the assortativity difference in time and modeled using a variant of independent cascade models.

Finally, to determine the source of assortativity in social networks, we described three statistical tests: the shuffle test, the edge-reversal test, and the randomization test. The first two can determine if influence is present in the data, and the last one can determine both influence and homophily. All tests require temporal data, where activation times and changes in attributes and links are available.

8.6 Bibliographic Notes

Indications of assortativity observed in the real world can be found in [62]. General reviews of the assortativity measuring methods discussed in this chapter can be found in [209, 212, 215].

Influence and homophily are extensively discussed in the social sciences literature (see [57, 192]). Interesting experiments in this area can be found in Milgram's seminal experiment on obedience to authority [194]. In his controversial study, Milgram showed many individuals, because of fear or their desire to appear cooperative, are willing to perform acts that are against their better judgment. He recruited participants in what seemingly looked like a learning experiment. Participants were told to administer increasingly severe electric shocks to another individual ("the learner") if he answered questions incorrectly. These shocks were from 15–450 volts (lethal level). In reality, the learner was an actor, a confederate of Milgram, and never received any shocks. However, the actor shouted loudly to demonstrate the painfulness of the shocks. Milgram found that 65% of participants in his experiments were willing to give lethal electric shocks up to 450 volts to the learner, after being given assurance statements such as "Although the shocks may be painful, there is no permanent tissue damage, so please go on," or given direct orders, such as "the experiment *requires* that you continue." Another study is the 32-year longitudinal study on the spread of obesity in social networks [54]. In this study, Christakis et al. analyzed a population of 12,067 individuals. The body mass index for these individuals was available from 1971–2003. They showed that an individual's likelihood of becoming obese over time increased by almost 60% if he or she had an obese friend. This likelihood decreased to around 40% for those with an obese sibling or spouse.

The analysis of influence and homophily is also an active topic in social media mining. For studies regarding influence and homophily online, refer to [297, 255, 62, 223, 300, 22]. The effect of influence and homophily on the social network has also been used for prediction purposes. For instance, Tang et al. [271] use the effect of homophily for trust prediction.

Modeling influence is challenging. For a review of threshold models, similar techniques, and challenges, see [107, 296, 108, 146].

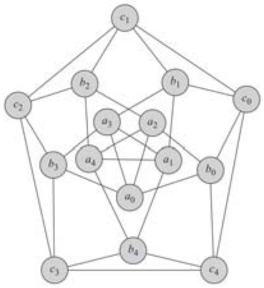
In addition to tests discussed for identifying influence or homophily, we refer readers to the works of Aral et al. [15] and Snijders et al. [261].

8.7 Exercises

1. State two common factors that explain why connected people are similar or vice versa.

Measuring Assortativity

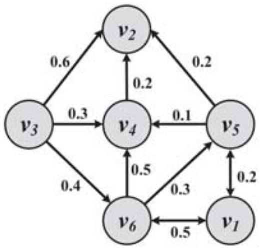
2. - What is the range $[\alpha_1, \alpha_2]$ for modularity Q values? Provide examples for both extreme values of the range, as well as cases where modularity becomes zero.
   - What are the limitations for modularity?
   - Compute modularity in the following graph. Assume that $\{a_i\}_{i=0}^4$ nodes are category a, $\{b_i\}_{i=0}^4$ nodes are category b, and $\{c_i\}_{i=0}^4$ nodes are category c.



Influence

3. Does the linear threshold model (LTM) converge? Why?

4. Follow the LTM procedure until convergence for the following graph. Assume all the thresholds are 0.5 and node v_1 is activated at time 0.



5. Discuss a methodology for identifying the influentials given multiple influence measures using the following scenario: on Twitter, one can use in-degree and number of retweets as two independent influence measures. How can you find the influentials by employing both measures?

Homophily

6. Design a measure for homophily that takes into account assortativity changes due to influence.

Distinguishing Influence and Homophily

7. What is a shuffle test designed for in the context of social influence? Describe how it is performed.

8. Describe how the edge-reversal test works. What is it used for?