Am I More Similar to My Followers or Followees? Analyzing Homophily Effect in Directed Social Networks

Mohammad Ali Abbasi, Reza Zafarani, Jiliang Tang, and Huan Liu
Computer Science and Engineering, Arizona State University
Tempe, AZ, USA
{Ali.Abbasi, Reza, Jiliang.Tang, Huan.Liu}@asu.edu

ABSTRACT

Homophily is the theory behind the formation of social ties between individuals with similar characteristics or interests. Based on homophily, in a social network it is expected to observe a higher degree of homogeneity among connected than disconnected people. Many researchers use this simple yet effective principal to infer users' missing information and interests based on the information provided by their neighbors. In a directed social network, the neighbors can be further divided into followers and followees. In this work, we investigate the homophily effect in a directed network. To explore the homophily effect in a directed network, we study if a user's personal preferences can be inferred from those of users connected to her (followers or followees). We also study the effectiveness of each of these two groups on prediction one's preferences.

Categories and Subject Descriptors

H.4 [**Database Applications**]: Data mining; D.2.8 [**Software Engineering**]: Metrics—*Complexity measures, Performance measures*

General Terms

Algorithms, Theory

Keywords

Social Media Mining; Preference Prediction; Relational Learning; Homophily

INTRODUCTION

Individuals extensively use online social networks to connect to other users, share information, express themselves, and benefit from the information provided by other users. In social networks, users often connect to those who have similar characteristics or similar interests. As a result, social networks are homogeneous with regards to many personal or behavioral characteristics [11]. *Homophily* is the tendency of similar individuals to form connections. The effect of this phenomena is a network in which connected users

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
HT'14, September 1–4, 2014, Santiago, Chile.
Copyright 2014 ACM 978-1-4503-2954-5/14/09 ...$15.00.
http://dx.doi.org/10.1145/2631775.2631828 .

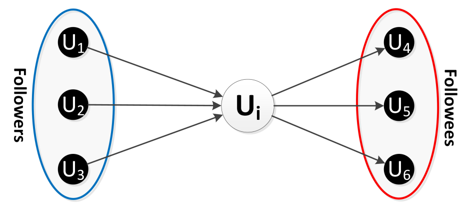


Figure 1: Neighbors in directed social networks can be grouped into *followers* and *followees*.

are more likely to share similar attributes and interests than disconnected users [18]. Homophily has its roots in undirected social networks, in which the two sides of the interaction are equally responsible to create and maintain the relation. Forming a real-world friendship is an example for this type of behavior. There is another type of connection that mostly appears in traditional mass media as well as online directed social networks. In this type of relation, only one party is responsible for the creation of the connection. Becoming a fan of an author or following a user on Twitter are examples of directed relation. In the example of an author and the group of her fans, the connection between the fans and the author is different from a regular friendship. The author has no control over these connections or does not even know many of her fans. Though the fans find themselves similar to the author, it cannot concluded that the author also will find herself similar to the her fans. In this example, the relation is formed and maintained solely by one of the parties involved in the relation. How can we measure the homophily effect in a directed social network? If the fans find themselves similar to the author, does this imply that the author will also reach the same conclusion?

A similar situation can be observed in many online social networks. In many networks, such as Facebook, the relation is bidirectional, where two connected users have to show their willingness for the relation to form. For instance, to form friendships on Facebook, one should initiate a friend request and the other user should accept it. However, in many social networks, the relations is directed. A directed connection, such as *following* on Twitter or *liking* on Facebook, is the result of only one user's action and there is often no need for the consent from the other user to get her involved in the relation.

In this work, we study homophily in directed social networks. To analyze homophily in directed networks, we study if a user's personal preferences can be inferred from her neighbors. Our goal is to determine which group (followers or followees) is more effective in inferring users' personal preferences. We conduct our experiments by using a set of more than 5 million Facebook fan pages to infer users' political orientation.

LITERATURE REVIEW

Predicting individuals' characteristics in the real-world has a long history. One often employs different types of information provided by the individuals such as their content, interactions with others, the products and services that they use, and the locations they visit [5, 1, 2] to infer users' characteristics and preferences. Similarly, in online social media, prediction techniques often use *content*, *user's interaction information*, or *network information* to infer user's profile attributes and preferences.

Content-based methods use user-generated content as a source to infer profile attributes information. By constructing features from the content and utilizing machine learning techniques such as Support Vector Machines (SVM) [16], Latent Dirichlet Allocation (LDA) [4], or boosted decision trees profile attributes can be predicted. *Interaction-based methods* utilize the interactions among users in online social networks to predict their profile attributes. Interactions include but are not limited to sharing contents of other users, commenting, retweeting, tagging, mentioning, or liking other users or their content [14]. *Network-based methods* use the network connections (links) to infer missing attributes. Recent studies show that it is possible to use information available from connected users in social networking sites to infer missing attributes and preferences with high accuracy [12, 13, 9, 10, 3, 6].

Tan et al. [17] used Twitter mention (@) data to construct a network and showed that users who mention each other in their tweets are more likely to hold similar opinions. Hu et al. [6] showed that connected people are more likely to be similar than randomly chosen disconnected people. Conover et al. [4] reported an accuracy of up to 95% when predicting users' political orientation by employing users' network information on Twitter. Mislove et al. [12] showed that it is possible to use only about 20% of the users providing attributes to infer the attributes for the rest of the network by an accuracy of over 80%. Carter et al. in [7] used the network structure and users' positions within a friendship network on Facebook to accurately predict users' sexual orientation. In a recent study by Kosinski et al. [8], the authors utilized Facebook data to show the degree to which relatively basic digital records of social media users' behavior can be used to accurately predict a wide range of personal attributes. They use Facebook *likes* to extract users' positive association with online content, such as photos, videos, Facebook pages of products, businesses, people, books, places, and websites.

In this study, we focus on network-based approaches, where we infer a user's missing attributes from the attribute information provided by other users in the network. Based on *homophily*, connected users more likely hold similar interests and attributes than those are not connected. This introduces a basic strategy to infer one's missing information based on the information of those connected users who revealed their attributes [15]. In its simplest form, one can use the user's neighbors to infer the user's missing information as well as her interests. This is a well-defined problem in undirected networks; however, it is not clear how one can use a user's neighbors to effectively infer missing information in directed networks.

HOMOPHILY-BASED PREDICTION OF USER'S PREFERENCES

In this section, we introduce our homophily-based approach for predicting user's preferences. We evaluate the predictive power of followers and followees for predicting users' profile attributes. We follow a two-step approach: first, we determine the level of homophily between users and their followers and users and their followees. Then we use followers and followees as independent sources to predict users' profile attributes.

Measuring Homophily

To measure homophily, one requires a method to compute homogeneity between users and their followees and followers. We employ a similar measure to the one outlined by Mislove et al. [12] to calculate the homophily among the users. Let a_i denote the value for attribute a for user u_i. We calculate the similarity among the user u_i and her neighbors $u_j \in N(u_i)$ on attribute a as

$$S_a = \frac{\sum_{u_j \in N(u_i)} \sigma(a_i, a_j)}{|N(u_i)|} \tag{1}$$

where $N(u_i)$ is a set of u_i's neighbors, and $\sigma(a_i, a_j)$ is the Kronecker delta function that returns 1 if the value of attribute a is equal for the two users and 0, otherwise.

$$\sigma(a_i, a_j) = \begin{cases} 1 & \text{if } a_i = a_j \\ 0 & \text{otherwise,} \end{cases}$$

$N(u_i)$ can be either u_i's followers or followees. For every user, we run the algorithm twice; first we use followers and then we use followees. In Equation 1, the value of S_a represents the fraction of the nodes with similar attribute values for the given attribute a.

To measure the statistical significance of S_a, we divide S_a by the expected value E_a when two users are chosen at random. Assume that attribute a can take k attribute values. Let A_i, denote the number of users that take the ith, $1 \leq i \leq k$ possible value for attribute a. Let $U = \sum_{i=1}^{k} A_i$ denote the total number of users. Then E_a can be computed as

$$E_a = \frac{\sum_{i=1}^{k} A_i(A_i - 1)}{|U|(|U| - 1)} \tag{2}$$

Let $H_a = \frac{S_i}{E_i}$ denote the degree of homophily between the user and her neighbors. When H_a is 1, there is no correlation between the attribute values. When it is less than 1, there is a negative correlation, and when it is greater than 1, it indicates a positive correlation between the attribute a's value of the user and the neighbors. Higher H_a indicates higher correlation between the attribute values of the user and that of the neighbors.

Predicting the Profile Attribute Values

The algorithm infers the given node's missing information by using the node's neighbors as the source of information. In this study, we use weighted majority vote to infer the user's profile attributes. To predict the value of attribute a for user u_i, we take the majority vote from u_i's neighbors regarding this attribute and assign the value with the highest number of votes.

EXPERIMENTS

As we described earlier, social media users decide whom to follow, however, they have no control on selecting their followers. Thus, we expect a higher degree of similarity between users and their followees than the users and their followers. We conduct two sets of experiments to evaluate the effect of homophily in directed social networks.

- *Observing the homophily*, to investigate the existence of homophily in directed networks, we measure and compare the similarity between users and their followees and users and their followers.

Table 1: Facebook Fan Pages Dataset Statistics

Total number of pages	5,856,000
Number of personal pages	764 K
Number of links	19,646,000
Revealed political orientation	25,129 (0.43%)

- *Investigating the prediction power or followers and followees*, we try to predict users' attributes, by using their followers and their followees and compare the results of two sources.

Dataset

In this study, we use *Facebook fan pages* to construct the directed social network. On Facebook, users can create regular user accounts or fan pages. Despite the regular user accounts on Facebook that form an undirected network, the fan pages' network is directed. To connect to a page, users (and pages) have to like the target page. This is similar to *following* behavior on Twitter. Each page can *like* or *be liked* by other pages and users. In the network, each page is a node and liking another page creates a link from the source node to the target node. There is no limit on the number of users that can like a Facebook fan page. The number of likes is a public property of the page and in our experiments is used to measure the popularity of the pages.

Data collection process.

Table 1 represents the statistics of our Facebook dataset. The dataset is collected by crawling Facebook through the site's public web interface. We start with a small set of seeds from the United States politicians, whose pages are publicly available on Facebook. We expand the set of seed nodes by following a *breadth first search (BFS)*. Thus, after we crawl all of the seed pages, we continue with the pages liked (followed) by the seeds, and this process is iterated until all the possible pages are collected. For every page, we collect the following publicly available attributes: *title*, *number of likes*, *political orientation*, *political party*, *category*, *gender*, and *list of liked pages*. Political orientation is an attribute with nominal values. The values and their distributions in our dataset are as follows: *Conservative* (23%), *Other* (20%), *Moderate* (19%), *Liberal* (18%), *Very Liberal* (7%), *Libertarian* (7%), *Very Conservative* (4%), and *Apathetic* (2%). In every step of the crawling task, we use *page category* to filter out pages that are not related to the US politics. Some of the relevant categories are *Politicians* and *Public Figures*. Every page has a category and most of the categories are chosen from a given list with predefined values. Frequency of categories follows a power-law distribution in which 19.8% of all pages belong to one category and 46% of them are in top 10 categories. Even when considering 90% of the pages, they belong to 160 categories. Table 2 shows a list of popular categories from our dataset. We use the category information to filter out pages do nor belong to persons. Therefore, in our final dataset that we use for experiments, every page belongs to a person such as a politician or a public figure.

Homophily in Directed Networks

Our goal in this experiment is to show whether a user is more similar to her followees or her followers and to verify if there is any significant difference between the two. We use the technique described in previous section to measure homophily and to evaluate the results. We use *political orientation* and *page category*, as the attributes for measuring homophily.

The experiments show that in more than 72% of cases, users have similar political orientation with their immediate neighbors, including followees and followers. In our dataset, the probability of holding the same political orientation for randomly chosen pairs of users is 25%. Next, we cluster the neighbors into two groups, including followers and followees. We observe a similarity of 73.5% between users and their followers, which is slightly higher than their 74% similarity with their followees. There is a slightly higher similarity between the user and her followees than the user and her followers. However the difference between these two results are not significant. There is a possibility that users' popularity influences our results. To investigate this possibility, we divide the users into two groups based on their popularity. A user is considered popular if she has more than 10,000 likes and non-popular if she has less than 1,000 likes. As we can see in Table 3, popular users are more politically aligned with their followees than non-popular users. In contrast, non-popular users are more likely to hold the same political orientation as their followers. One explanation for this observation is that popular and non-popular users' liking behavior could be different. Popular users, comparing to the number of followers, often have much smaller number of followees that are chosen very carefully. Therefore, we expect to observe a higher degree of similarity between a popular user and her followees. Popular users have too many followers and these followers might have reasons other than holding the same political orientation for following the popular individual. Non-popular users, on the other hand, are more eager to attract more followers, therefore they follow other users hoping that these users would follow them back. These users, usually have diverse attribute which lead to poor prediction accuracy. Therefore, non-popular users are less likely to share similar attributes with their followees. On the other hand, the small set of followers of non-popular users should have a good reason to follow them. Therefore, there is a higher chance that a non-popular user and her followers share similar interests or attributes, such as a political orientation.

Table 2: Popular Categories of Facebook Pages and their Popularity Level in our Dataset

Rank	Category	Fraction of pages
1	Community	16.8%
2	Musician/Band	7.4%
3	Non-Profit Organizations	4.1%
4	Public figure	3.8%
39	Politician	0.4%
49	Political Organization	0.3%
81	Political Party	0.2%

Table 3: Political Orientation Consistency between Users, Their Followees, and Their Followers with Respect to Different Levels of User Popularity

Neighbors	All	$\leq 1K$	$\geq 10K$	≤ 100	$\geq 1M$
Followees	74%	75%	75%	73%	73%
Followers	73.5%	73%	74%	76%	74%
Fe + Fr	72%	72%	73%	73%	72%

Page Category.

We run the same set of experiments for measuring homophily, but instead of the *political orientation* attribute, we use the *page category*. Table 4 shows the results. On average, 35% of the connected users belong to a similar category. Users in 39% of cases have a similar category with their followees and in 37% of the cases with their followers, which in both cases is higher than using a combination of followers and followees.

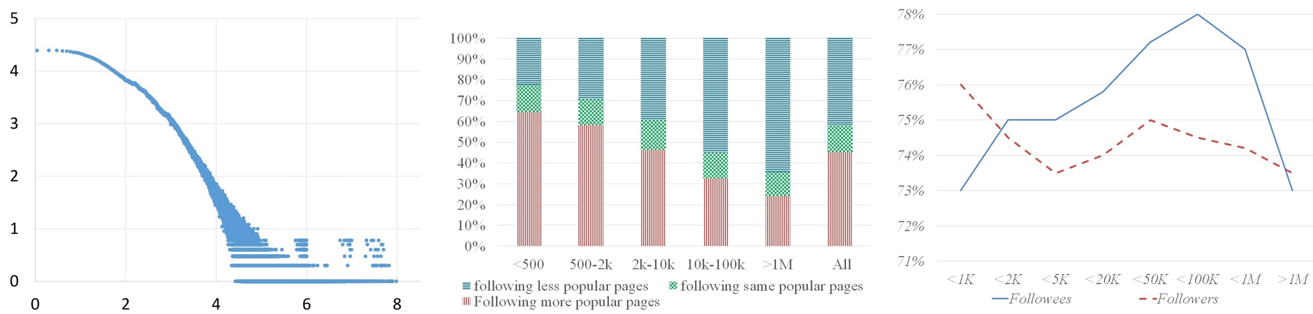


Figure 2: a) User popularity distribution, x-axis is popularity (logarithm of number of page likes) and y-axis is the frequency of pages holding that popularity. b) More than 62% of Facebook pages like pages that are more popular than or equally popular as the page. c) The effect of pages' popularity on liking similar pages. The graph shows the similarity between a page's political orientation with her followees and followers. X-axis is the page's number of likes

The Effect of Popularity on Homophily

Popularity is an attribute that is correlated with the page's number of likes. As popularity follows a power-law distribution, we compute the logarithm of the number of user's likes, $\log(\text{likes}(u_i))$, to discretize the attribute into 8 categories. Figure 2.(a) shows the popularity distribution in our dataset. To evaluate the effect of page popularity on users' following behavior, we measure the relative popularity of each page and popularity of her followers and popularity of her followees. Results indicate that in 49.5% of cases, users like more popular users. In 12.5% of cases, users like users with the same popularity level, and in 38% of cases users like less popular users. This result matches with our expectation that users usually follow those who are more popular than themselves. Figure 2.(b) shows this behavior with respect to different popularity levels. As we can see in this figure, 24% of extremely popular users like users that are not as popular as themselves. Users with more than 2,000 and less than 10,000 likes are the most balanced group of users with respect to following and being followed by users with the same popularity level.

To evaluate the effect of users' popularity on their following (liking) behavior, we measure the homophily of each group of users with respect to their popularity level and plot the results in Figure 2.(c). As we can see in this graph, overall, followees are a better match with users than their followers, although there are some exceptions. Users with less than 100 likes highly match with those who liked them. When the popularity increases, we observe a higher homophily effect between users and those they like (follow). The maximum homophily effect belongs to users with about 100K likes. Beyond that, the trend changes and the curve touches its minimum level of similarity, which belongs to celebrities. Celebrities, usually have a non-uniform liking behavior. They follow users from different categories and different popularities, which decreases the similarity between the user and her followees. The same effect occurs with those who follow celebrities, as a celebrity has followers from a variety of categories and interests, which decreases the similarity between the celebrity user and her followers.

Table 4: Category Consistency between Users, Their Followees, and Their Followers with respect to Different Levels of User Popularity

Neighbors	All	$\leq 1K$	$\geq 10K$	≤ 100	$\geq 1M$
Followees	39%	44%	35%	30%	30%
Followers	37%	39%	34%	33%	26%
Fe + Fr	35%	40%	33%	31%	26%

Neighbors Diversity

In this section, we investigate the effect of neighbor diversity on homophily. We use entropy to measure the diversity among followers and followees as follows,

$$e_i = -\sum_k P(A_k) log P(A_k) \tag{3}$$

where A_i is the number of users that take the ith, $1 \leq i \leq k$ possible value for attribute a and e_i is the entropy of user u_i's neighbors with respect to attribute a. Higher entropy indicates the higher diversity among one's neighbors. We calculate the entropy for followers e_{i_r} and followees e_{i_e}. We summarize the results in Figure 3 considering the following possible scenario: $e_{i_r} \approx e_{i_e}$, $e_{i_r} > e_{i_e}$, or $e_{i_r} < e_{i_e}$.

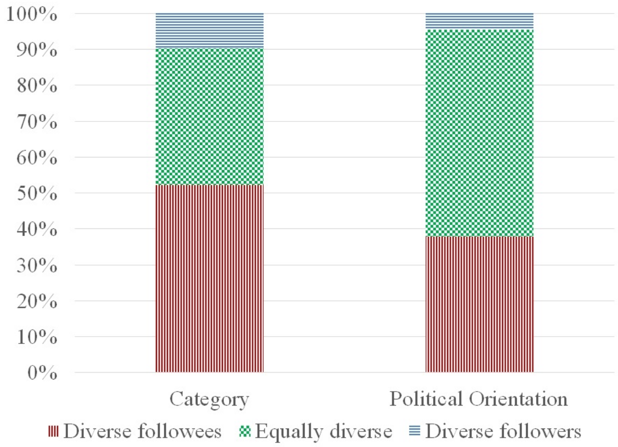
Each bar in Figure 3 shows three values. The blue bar shows the percentage of users who have more diverse followees than followers, the red bar shows the percentage of users who have as diverse followers as followees, and the green bar the percentage of the users who have more diverse followers than followees. For both of the attributes, *political orientation* and *page category*, followees are more diverse than followers. Looking at this problem from a user popularity point of view, users with less than 1,000 likes follow the most diverse group of users. In contrast, popular users and celebrities hold the smallest percentage of diverse followees. Diversity among the follower and followees is a measure that can be used to decide which source should be used to infer users' missing information.

Neighbors' Prediction Power

In this section, we investigate the neighbors' prediction power. We use followees, followers, and the combination of followees and followers to predict users' missing information. As previously mentioned, we use weighted majority vote to infer users' missing information. Similar to the previous section, we use followees and followers to predict users's *political orientation* and *category* .

Predicting Political Orientation

In these sets of experiments, we used immediate neighbors to predict users' missing information. The results show that if we use all of the users' neighbors, including followees and followers, by using majority vote algorithm, we can achieve 75% accuracy in predicting users' political orientation. If we limit the neighbors to



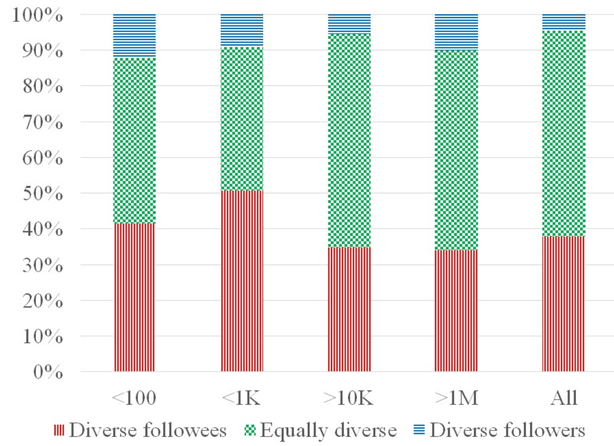


Figure 3: Neighbor diversity among followers and followees. For both attributes, page category and political orientation, followees are more diverse than followers. The figure on right shows the political orientation diversity for pages with different popularity level.

Table 5: Predicting political Orientation using Users' neighbors

Neighbors	All	$\leq 1K$	$\geq 10K$	≤ 100	$\geq 1M$
Followees	77%	78%	76%	78%	78%
Followers	73%	72%	72%	73%	72%
Fe + Fr	74%	74%	74%	74%	74%

Table 6: Page Category Prediction using Users' Neighbors

Neighbors	All	$\leq 1K$	$\geq 10K$	≤ 100	$\geq 1M$
Followees	45%	47%	40%	29%	31%
Followers	43%	36%	38%	32%	26%
Fe + Fr	43%	36%	39%	33%	28%

only the followees, the accuracy increases to 77%. By using one's followers to predict her information we are able to achieve 73% accuracy, which is less than followees and a combination of followees and followers. Table 5 shows the detailed results with respect to different levels of user popularity. The results show that in all different experiments, followees are better sources to predict users' political orientation. Though followers are not as good as followees, they can correctly predict political orientation in more than 73% of cases. Similar to the results from Section using a combination of followers and followees does not improve the accuracy compared to just using the followees.

Predicting Category

Similar to predicting political orientation, we used neighbors to predict users' category. Using all neighbors generates 43% accuracy which is less than followees with 45% accuracy and is similar to followers with 43% accuracy. Prediction results with respect to different levels of users' are reported in Table 6.

CONCLUSION

Our goal in this paper was to study homophily in directed social networks. We investigated whether one can use the neighbors in directed networks to infer users' preferences. We use a dataset of 5 million Facebook fan pages and form a directed network to conduct experiments. We divide every users' neighbors into followers and followees, and use them to infer users' personal preferences. The experiments revealed one's followees can be used to predict her preferences with 74% accuracy. With a similar setting followers predict users' preference with 73.5% accuracy. The results show the effectiveness of both followers and followees on predicting one's preferences. Previous work on inferring missing attributes in social networks show that it is possible to predict users' personal preferences by using their own online behavior. In this study, we show that not only users' own online behavior, but also users' neighbors' behavior can be used to reveal users' attributes and preferences. Our findings raise the awareness of users over the dangers of having their privacy violated by being able to predict their preferences using individuals that follow them.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their useful comments. This research is, in part, sponsored by the Office of Naval Research grants N000141110527 and N000141410095.

REFERENCES

[1] M. A. Abbasi, S.-K. Chai, H. Liu, and K. Sagoo. Real-world behavior analysis through a social media lens. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 18–26. Springer, 2012.

[2] M. A. Abbasi and H. Liu. Measuring user credibility in social media. In *Social Computing, Behavioral-Cultural Modeling and Prediction*, pages 441–448. Springer, 2013.

[3] A. Chaabane, G. Acs, M. A. Kaafar, et al. You are what you like! information leakage through users' interests. In *Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS)*, 2012.

[4] M. D. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer. Predicting the political alignment of twitter users. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 192–199. IEEE, 2011.

[5] S. D. Gosling, S. J. Ko, T. Mannarelli, and M. E. Morris. A room with a cue: personality judgments based on offices and bedrooms. *Journal of personality and social psychology*, 82(3):379, 2002.

[6] J. Hu, H.-J. Zeng, H. Li, C. Niu, and Z. Chen. Demographic prediction based on user's browsing behavior. In

Proceedings of the 16th international conference on World Wide Web, pages 151–160. ACM, 2007.

[7] C. Jernigan and B. F. Mistree. Gaydar: Facebook friendships expose sexual orientation. *First Monday*, 14(10), 2009.

[8] M. Kosinski, D. Stillwell, and T. Graepel. Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences*, 110(15):5802–5805, 2013.

[9] R. Li, S. Wang, H. Deng, R. Wang, and K. C.-C. Chang. Towards social user profiling: unified and discriminative influence model for inferring home locations. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1023–1031. ACM, 2012.

[10] B. Marcus, F. Machilek, and A. Schütz. Personality in cyberspace: personal web sites as media for personality expressions and impressions. *Journal of Personality and Social Psychology*, 90(6):1014, 2006.

[11] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[12] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel. You are who you know: inferring user profiles in online social networks. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 251–260. ACM, 2010.

[13] D. Murray and K. Durrell. Inferring demographic attributes of anonymous internet users. In *Web Usage Analysis and User Profiling*, pages 7–20. Springer, 2000.

[14] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft. Our twitter profiles, our selves: Predicting personality with twitter. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 180–185. IEEE, 2011.

[15] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowcroft. The personality of popular facebook users. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*, pages 955–964. ACM, 2012.

[16] D. Rao and D. Yarowsky. Detecting latent user properties in social media. In *Proc. of the NIPS MLSN Workshop*, 2010.

[17] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level sentiment analysis incorporating social networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1397–1405. ACM, 2011.

[18] R. Zafarani, M. A. Abbasi, and H. Liu. *Social Media Mining: An Introduction*. Cambridge University Press, 2014.