How Much do Your Friends Tell About You?

Reconstructing Private Information from the Friendship Graph

Norbert Blenn Christian Doerr Nasireddin Shadravan Piet Van Mieghem

Faculty of Electrical Engineering, Mathematics and Computer Science
Delft University of Technology, P.O. Box 5031, 2600 GA Delft, The Netherlands
N.Blenn@tudelft.nl, C.Doerr@tudelft.nl, N.Shadravan@student.tudelft.nl, P.F.A.VanMieghem@tudelft.nl

Abstract

After the early land rush and fast exponential growth of online social networking platforms, concerns about how data placed in online social networks may be exploited and abused have begun to appear among mainstream users. Social networking sites have responded to these new public sentiments by introducing privacy filters to their site, allowing users to specify which aspects of their profile are visible to whom. In this paper, we demonstrate that such an approach to privacy and informational self-determination is largely futile: as we form social relations and build networks with those alike us, much of who we are and what we do can be reconstructed from unhidden parts of the social graph.

Categories and Subject Descriptors K.4.1 [*COMPUTERS AND SOCIETY*]: Public Policy Issues—Privacy

General Terms Security, Online Social Network

Keywords Privacy, OSN, Friendship Graph

1. Introduction

The recent boom of social networking platforms has lead to a dramatic shift in how people behave, spend their time and interact with others. The wealth of information registered users and visitors voluntarily place, curate and maintain within these platforms in combination with their enormous market reach has however also enabled a wide set of new applications beyond the initial usage propositions: Activities of users and their interactions with their friends are now analyzed to obtain personal profiles, which can be used for marketing activities, but also help companies determine whether a customer can be deemed "influential" and should consequently receive a better treatment than others [13]. Information on relationships, personal habits and interests can be taken into account when assessing risks and rates when applying for health insurance [16], and face recognition performed on photos stored in online social media allows the re-identifications of persons in other contexts, such as identifying passerby's in camera recordings to deliver targeted billboard advertisements [17].

As such technologies are developed and applied, concerns about the privacy of one's personal data are increasingly gaining track. Indeed, privacy filter usage has become a mainstream practice: in case of the largest national social network site in the Netherlands, hyves.nl, 63% of the users have by now enabled privacy settings in their profile making their details invisible to the general public.

In this paper, we demonstrate to what extent and at which accuracy level personal information can actually be reconstructed from a social network's friendship graphs. The underlying justification our approach is driven by is the sociopsychological hypothesis, which was empirically verified for digg.com [5] and facebook.com [14], that users form social ties with those around them who are similar in socio-economic status, interests and opinions [15]. In consequence, knowing a user's friends can therefore to a large degree tell us the individual tastes and choices of a social network user even when his profile page is hidden.

The degree to which this technique can be successfully applied varies with the overall embedding of a particular user in the social graph as well as other attributes, such as the user's personal characteristics, the overall diversity of the direct friends or the degree to which the friends are making use of privacy settings themselves.

The remainder of this paper is structured as follows: Section 2 overviews previous work on privacy in social networking sites, section 3 outlines the platform Hyves.nl and the data acquisition used for this study. Section 4 demonstrates a study how hidden profile information can be extrapolated from the social friendship graph. Section 5 summarizes our findings and gives an outlook on future work.

5th Workshop on Social Network Systems EuroSys '12 April 10, Bern.

2. Related Work

Two major approaches, active and passive, are possible to access private information. Active approaches try to obtain data by directly attacking a particular user using fake profile information [2], surveys or third party applications that access the users profile in the OSN. In this paper we will investigate passive approaches which are based on statistical analyses of users and the friendship network. These passive approaches may be based on the profile information a user specifies, tracking the friendship network through third-party applications, or the combination of different data sources.

Gross and Acquisti [7] analyzed patterns of information revelation in OSNs and privacy implications in the "early" stage of Facebook. An amazingly high number of 89% of users in their dataset provided their real name. Other attributes like phone number, birthday, home town, address etc. were also given by the majority of the users. Different techniques to infer private information like reidentification of users by analyzing the postal code and their birthday are presented. Face re-identification to identify users on different sites or even identity theft of the users social security number was shown to be feasible.

The role of third party sites in tracking users of OSNs and obtaining private information is investigated by Krishnamurthy and Wills [9, 10]. In most cases, a user has no possibility to control all applications that track profile data. Users are not aware which data is accessed by them and what the different services do with this data.

Because of the knowledge about friendships in OSN and the fact that those relations are mostly built between individuals having similar interests it is still possible to infer private attributes of a user from his friends even if the user has a profile which is not visible to everyone. McPherson et al. [11] discussed "homophily" as a concept that limits individuals to connect only to others having similar attributes. The strongest divisions are based on race and ethnicity followed by age, religion, education, occupation and gender. Hence, ties between non-similar users are either not constructed or dissolve at a higher rate. This leads to social niches in the social space.

He et al. [8] constructed a Bayesian network assuming that direct neighbors have a higher overlap than users multiple hops away. It is shown that privacy can be indirectly inferred via social relations and mathematically over multiple hops. He et al. use an influence strength which is defined as the conditional probability ($P(A|B)$) that user A has a attribute given a friend (B) has the same attribute.

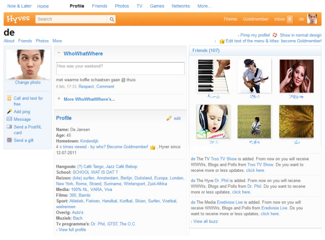
By using friendship information and group attendance information, Zheleva and Getoor [19] showed for different OSNs that it is possible to infer private attributes using group and friendship information.

Mislove et al. [14] claim that "you are who you know" because automatic community detection for multiple attributes of the users led them infer private attributes with an accuracy of 80% inside those communities. This approach needs the knowledge of the topology of the social network in order to detect communities. Because of the dynamic nature of OSNs, standard crawling techniques take rather long to obtain the whole network, it is thus unfeasible for attackers to first crawl the network in order to detect communities.

3. Hyves.nl

Hyves.nl is the largest Dutch Online Social Network founded in 2004 containing nearly 10.6 million accounts in 2011. Given the total population of the Netherlands (ca. 16.5 million), a large fraction of the inhabitants are registered. However, the total number of user accounts of Hyves.nl includes duplicates and orphan accounts as well as commercial pages.

We obtained our dataset by screen scraping Hyves.nl using multiple parallel breadth first searches. Our dataset contains 2,971,261 user profiles. Out of those roughly one third are public viewable profiles. On a profile page, users have to define a username and a real name and they may provide birthday, age, hometown, relationship status, living situation, address, phone number and their email address. An example of a typical Hyves.nl profile page is shown in fig. 1.



met waarme koffie schaatsen gaan @ thuis
6 feb, 17:33, Respect, Comment

Name: De Jansen
Age: 45
Hometown: Kinderdijk
4 x times viewed - by who? Become Goldmember! , Hyver since 12-07-2011

Hangouts: (?) Café Tango, Jazz Café Bebop
School: SCHOOL WAT IS DAT ?
Reizen: (kite) surfen, Amsterdam, Berlijn, Duitsland, Europa, Londen, New York, Rome, Strand, Suriname, Wintersport, Zuid-Afrika
Media: 100% NL, VARA, Viva
Films: 300, Bambi
Sport: Atletiek, Fietsen, Handbal, Korfbal, Skien, Surfen, Voetbal, wielrennen
Overig: Auto's
Muziek: Bach
Tv programma's: Dr. Phil, GTST, The O.C.
› View full profile

de The TV Tros TV Show is added. From now on you will receive WWWs, Blogs and Polls from Tros TV Show. Do you want to receive more or less updates, click here.

de The Hyve Dr. Phil is added. From now on you will receive WWWs, Blogs and Polls from Dr. Phil. Do you want to receive more or less updates, click here.

de The Media Eredivisie Live is added. From now on you will receive WWWs, Blogs and Polls from Eredivisie Live. Do you want to receive more or less updates, click here.

Figure 1. Screenshot of a users main profile page of Hyves.nl

Additionally, users may join a large selection of groups. Those groups could be real world communities like sport clubs, schools or companies, famous people, bars and restaurants, books, movies etc. Groups are organized in 19 topics namely: brands, hangouts, school, college, club, company, TV shows, books, food, film, gadgets, games, famous people, media, music, traveling, sport, TV programs and others. It is possible to join groups without invitation and a user may create a new group. Groups are displayed on the user's profile page ordered by their topic.

Every group has its own page, listing all members of this group, additional information like events, addresses or opening times. Friendship relations are set up by sending a friendship request via Hyves.nl to a user. If the request is accepted the two users are mutual friends. The average number of friends in our dataset equals 127. Users may also upload photos and tag people in these photos. We also crawled 446,868 images and people tagged in them resulting in 624,478 user names 1,311,423 relations.

In terms of privacy control, Hyves.nl allows a user to change privacy settings to display each attribute to the public, viewable for everyone registered at Hyves.nl, friends of friends or only friends. Nearly one third of all profiles we collected, are publicly viewable, which means that the real name, groups, age, hometown and the list of friends is displayed. If a user has a private profile page, the real name, if entered by the user, is still displayed.

4. Reconstructing Users' Profiles

In order to infer private attributes of a user who has his profile page set to private, we use statistical methods based on different sources of information. For some OSN's like Hyves.nl, StudiVZ.de, Skyrock.com or Vkontakte.ru most users belong to the population of one country. Therefore, combining information from census bureaus of these countries constitutes a straightforward way of inferring attributes like the age, name, phone number or relationship of the user. We used association rule learning, trained on the dataset of publicly available information in order to reveal relations between user's attributes and groups a profile page lists. A third approach is based on the theory of "birds of a feather flock together" describing that friends have similar interests.

In general, the characteristics of a user can be classified into two groups: intrinsic attributes (such as name, age, city and the gender) and communities (school, college, university, company, sport club or interests).

If a user has a private profile page it is still possible to uncover friendship relations because they are bidirectional. Therefore these friendships are listed on profile pages of friends having a publicly viewable profile page. Based on the average number of friends this indicates that on average every user has 42 friends having a publicly viewable profile page. As stated in Bonneau et al. [3] "eight friends are enough" to reveal the whole network of users of a OSN. We will show that a similarly small number of friends is sufficient to correctly infer most of a user's characteristics.

4.1 "Birds of a Feather"

As described by McPherson et al. [11], friends tend to have similar interests because they know each other, live close to each other, met physically at places where they follow their hobbies or at places where they work together. Friendships in Online Social Networks do not necessarily follow this scheme as one may also create friendship relations towards users without knowing them in person. The hypothesis that personal preferences limit the possible number of users, still holds as online friendships are based on common interests as shown in [5, 11, 14, 15].

4.1.1 Age

The age distribution of users in Hyves.nl shows an oversampling of young persons when comparing to the age of the Dutch population as shown in fig. 2.

One assumption is that a user is as old as most of his friends. Hence, for every user in our dataset providing his age we used the most frequent age of his friends as an estimator. The results are shown in figure 3 given by the difference between actual age of a user to the mode of the friend's ages.

As indicated by multiple traces (different markers, and colors) in fig. 3, the probability that most friends have the same age as a user is depending on the age group the user is in. The highest accuracy of this method (prediction rate) is found for the group of 16 to 20 year old users where 61% of friends of a user have exactly the same age as the user. When allowing up to ±1 year of difference the probability to predict the correct age of a user, by using the age of most friends, increases to 77%. This prediction probability decreases for older age groups.

A reason for this high age overlap might be based on the fact that friendships in the group of 10 to 20 year old users are created in schools where the students are in the same class. Later in life, colleagues and friends are not exactly the same age anymore. Another explanation for this trend is the decreasing average number of friends: In the group of 16-20 year old users 81, users at the age of 46-50 years have on average 14 friends. We also found a peak in the prediction error for users between 35 and 45 years which is around 20 years, indicating that quite a number of parents are befriended with their children.

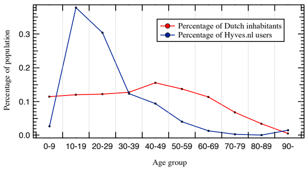


Figure 2. Comparison of the age of Hyves.nl users (blue) to the population of the Netherlands (red)

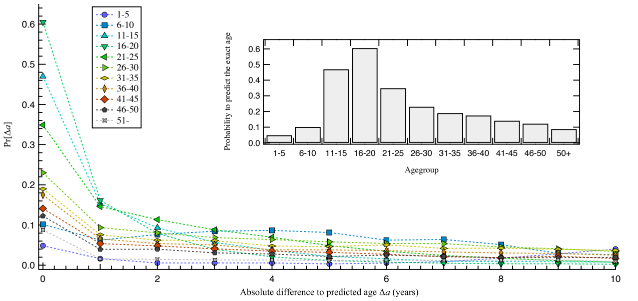


Figure 3. Prediction of the age of a user using mode of his friends age for different age groups. Inset: Probability to correctly guess the age of a user.

4.1.2 Location Determination

Another intrinsic value of a user is the hometown. The simplest assumption is that most friends of a user live in the same city the user lives in. Actually, based on our dataset 67% of a particular user's friends who provided their hometown live in the same city as the user, and 91% of all friends live within 50 km of the users location.

As already mentioned even private profile pages list the real name of a user. By accessing a database containing the geographical distribution of names, for example phone books, this information can be combined to estimate the area or even the city of a user. Some census bureaus or universities also collect data on name distributions. In the Netherlands the Meertens Institute [1] lists how often and in which municipality a particular family name or first name occurs. Because no name is evenly geographically distributed, one can use the name of a user to guess his location. Fig. 4 shows the geographical distribution and the popularity of "Tim" as first name during the last 131 years. A similar geographical distribution can be obtained for family names. Comparing the geographical distributions of first and family name provides a rough estimate for the city of a user.

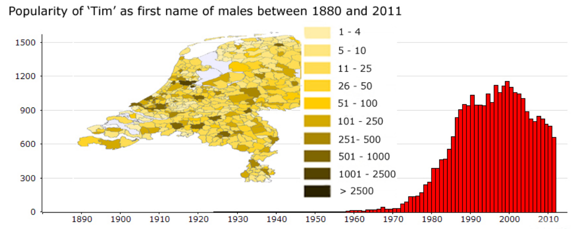


Figure 4. The popularity of the first name Tim in the Netherlands. Inset: Geographical distribution of the name.

Given the area distribution of names in fig. 4 and the distribution of the surname, we deduce that a user called "Tim Janssen"[1] lives in Amsterdam with a probability of 32%, the Hague with a probability of 19% or Utrecht with a probability of 17% etc.

Out of the 19 topic groups in Hyves.nl, some further strengthen location estimation. Topics like hangouts, schools, colleges, clubs, companies, food and sport contain implicit location information. Assuming that people like to visit bars, restaurants, spot clubs in the same city they live and work enables us to infer the city from these groups.

By using Bayesian analysis [18] we calculated the probability a user has joined a specific group given he lives in a specific city. In this way we compared how many users attend a group in different cities. If the distribution shows no significant peaks (larger than 1 standard deviation), this means that the users in this particular group are homogeneously distributed in the Netherlands. An over-representation of a particular group in a city however is a good indicator that this group can be used to infer a users city. We identified 13,512 groups that can be used to predict the residence of a user.

By analyzing all of those 13,512 groups with more than 5 users, we found that on average 64% of the members indeed live in the same city. This does not imply that the other 36% reside in different cities as some users simply do not provide their home town. When assuming that users, who did not enter a city in their profile, would live in the same city as most users of this group, the average predictability increases to 86%.

4.1.3 Different Tastes in Age Groups

Groups do not only reveal location information but also information about the age of user. For example musical interests have a strong relation to the age of a particular user. We depict in figure 5 exemplarily the age of users who like different singers or music bands. Conversely these correlations suggest that the specified age of most users in our dataset is accurate. We found strong relations between interests and the age of a user for movies, music types and game consoles.

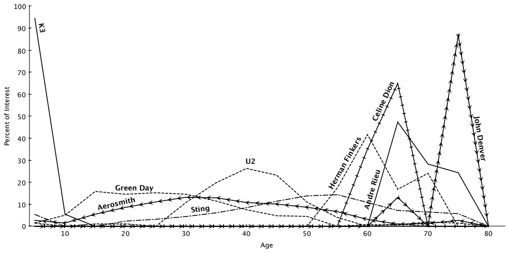


Figure 5. Probability users have a specific taste in music to the age of a user

Every user joined on average 26.6 groups. Homophily suggests that friends have similar tastes which should result in a high overlap of group memberships between a user and his friends. Figure 6 depicts the probability the profile page of a user's friend lists the same groups as the user. If all friends are taken into consideration only a small overlap (red points) can be found. When searching for the highest overlap of groups a user has with at least one friend we found that most users have at least one friend who joined nearly the same groups as the user (blue points). The difference in groups a user is a member of compared to his friends, can be seen as a similarity measurement between users. Based on this metric, only a fraction of all friends are close friends, whereas a high number of acquaintances appear in a user's friendship network. The fact that only a few friends in the friendship graph are close friends is also described by [6] as the weak and strong ties and analyzed by [12]. Thus a way of identifying close friends could be by analyzing the information if the users are tagged on the same image. This would imply that the users physically know each other and the probability they are close friends increases.

4.2 Association Rules

Association rule learning is a popular method used in data mining in order to discover relations between attributes in datasets. Often utilized for market basket analysis the input dataset for association rule learning contains an item set of things a person has bought. A typical rule created out of a supermarket dataset could therefore be the following: If noodles and cheese are bought then the customer will also

[1] Name is randomly chosen.

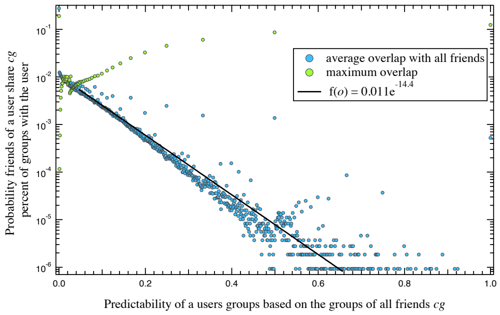


Figure 6. The percentage to which extend lists of groups between friends are similar. Red, the average overlap between all friends of a user. Blue, the maximal overlap with at least one of the friends.

buy bolognese sauce with a confidence of α percent where all products appear in β percent (support) of all purchases. The confidence α corresponds to the fraction of the support of all items in the rule to the support of the requisites. The naive way of calculating simple co-occurrences would result in a very large co-occurrence matrix because of ca. 1.1 million groups in our dataset. Given the groups of all users as input, association rule learning will still calculate rules in a reasonable time, for a given minimum support and confidence.

We used an implementation called apriori [4] to calculate association rules with a given minimum support of 0.1% and a minimum confidence of 50%. The exact number of groups in our dataset was 1115558. The support of 0.1% means that 1116 user profiles should list a group in order to include the group in the rule. The calculated rules had a maximum length of 4 resulting that at most 3 groups lead to a consequence. Longer rules are clear subsets of shorter ones having a higher confidence but smaller support. An example for such a rule is the following. We found that users that are interested in the soccer club "Ajax Amsterdam" are also interested in the "Amsterdam Arena" with a support of 0.203% and a confidence of 58%. But if a user is interested in "Ajax Amsterdam" and "Adidas" he is more likely to be interested in the "Amsterdam Arena" with a confidence if 83% but the rule has a support of only 0.113%.

Because it is possible to set the privacy settings for groups to only show groups out of selected topics, association rules learning helps to infer others. By knowing only a few groups of a user it is possible to directly apply a rule with a high confidence to infer other groups of the user.

The same holds for the earlier mentioned age prediction as shown in figure 5 based on different groups. For example the probability to be 11 years old if a user likes the movie "Finding Nemo" is 70%. If we know that the user additionally likes "Happy Feet" this probability increases to 87% as the rule gets more specific.

Interestingly the given example of soccer fans already depicts that group predictions work across different topics (sport to brands to hangouts). A graph of rules illustrating these connections is shown in fig. 7. The colors indicate different networks where the nodes are groups with their corresponding id's in our dataset.

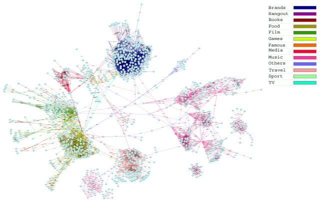


Figure 7. Graph of association rules. Nodes are groups labeled by their group id's. A link is drawn if a rule exist containing both groups. The links are labeled by the confidence value of the rule.

It is visible that most rules are between groups of the same topic (same color). For every topic there seems to be a few hubs standing for the largest groups in this particular topic that can be predicted by multiple other smaller groups.

As association rule learning seems to be a good solution to obtain global information about group predictions it is not an user-centric method. This means it is not possible to observe effects of the underlying topology of the friendship network. Therefore we calculated the "predictability" of a user using all rules. This predictability is defined by two values. One is the number of groups that can be inferred using all rules whereas the second is given by the average confidence of rules applied to all groups of a user. The latter gives insights in the "predictability" of this user. For every user we looked at all of his groups and calculated the average confidence of all rules that can be applied to its groups. Figure 8 depicts the predictability versus the fraction of predicted groups. A positive Pearson correlation can be found with a value of 0.537.

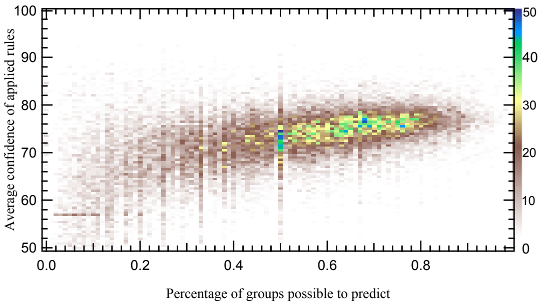


Figure 8. Joint 2D histogram of the percentage of groups that can be revealed using association rules versus the average confidence of the applied rules. The color indicates the number of users the rules apply to.

The Pearson correlation of the age of a user towards its predictability is slightly negative with -0.15 which in turn is based on the fact that the number of users in our dataset decreases for older users. As previously shown groups may have a certain dependency on the age which means that the groups, older people follow do not reach the required minimum size of 1116 users to be included as a result of association rule learning.

By comparing the predictability of a user with the predictability of the friends of this user we found a positive correlation value of 0.29. This indicates that users with a high predictability are connected to others having also a high predictability. By correlating the number of friends that have a publicly viewable profile to the predictability we found no significant relation, which means that a small number of friends having an open profile are already enough to guess the groups a user attends.

5. Conclusion

We showed that the friends of young users are in most cases as old as the user itself. In contrast if the age of the friends has a large variation the user is most probably an older one. However by looking at interests of users, or the friends interest the predictability of older users can be raised again to a percentage of 78%. We also showed that the number of friends needed to infer private attributes is relatively small.

Close friends have a very high overlap in terms of their groups they like and the city they live in. Most of the acquaintances of a user have dispersed attributes and only a small overlap. By identifying these few close friends a high prediction accuracy can be reached.

We showed in a case study how to infer private attributes of a user in an Online Social Network. By using statistical analysis were we were able to calculate rules that allowed us to reconstruct most of the interests of a user even if he has a private profile which is not viewable by everyone. We connected this information with information of friends of a user and showed that basic attributes like age and hometown can be derived with a very high accuracy. We also used data from statistical institutes to further increase the prediction rate.

Our findings lead to the conclusion that the common practice in privacy regulation is not practical at all. For most users in our dataset we were able to estimate private attributes.

Acknowledgments

This work was partly funded by the Trans-sector Research Academy for complex Networks and Services (TRANS).

References

[1] The meertens institute. Website, 2012. `http://www.meertens.knaw.nl/cms/en/meertens-institute`.

[2] L. Bilge, T. Strufe, D. Balzarotti, and E. Kirda. All your contacts are belong to us: Automated identity theft attacks on social networks. In *18th International World Wide Web Conference*, pages 551–551, April 2009.

[3] J. Bonneau, J. Anderson, R. Anderson, and F. Stajano. Eight Friends are Enough: Social Graph Approximation via Public Listings. In *Proceedings of Second ACM Workshop on Social Network Systems*, Mar. 2009.

[4] C. B. Department and C. Borgelt. Efficient Implementations of Apriori and Eclat, 2003.

[5] C. Doerr, S. Tang, N. Blenn, and P. Van Mieghem. Are friends overrated? a study for the social news aggregator digg.com. In *NETWORKING 2011, Part II*, Lecture Notes in Computer Science 6641, pages 314–327. IFIP International Federation for Information Processing, 2011.

[6] M. S. Granovetter. The strength of weak ties. *American Journal of Sociology*, 78(6):pp. 1360–1380, 1973. ISSN 00029602.

[7] R. Gross and A. Acquisti. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*, WPES '05, pages 71–80, New York, NY, USA, 2005. ACM.

[8] J. He, W. Chu, and Z. Liu. Inferring Privacy Information from Social Networks. In *Intelligence and Security Informatics*, volume 3975 of *Lecture Notes in Computer Science*, pages 154–165. Springer-Verlag, Berlin/Heidelberg, 2006.

[9] B. Krishnamurthy and C. E. Wills. Characterizing privacy in online social networks. In *Proceedings of the first workshop on Online social networks*, 2008.

[10] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. *SIGCOMM Comput. Commun. Rev.*, 40:112–117, Jan. 2010. ISSN 0146-4833.

[11] M. McPherson, L. S. Lovin, and J. M. Cook. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, 27(1):415–444, 2001. doi: 10.1146/annurev.soc.27.1.415.

[12] M. Mcpherson, L. Smith-lovin, and M. E. Brashears. Social isolation in america: Changes in core discussion networks over two decades. *American Sociological Review*, 71:353–375, 2006.

[13] A. Miller. Untangling the social web. *The Economist*, September 2 2010.

[14] 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 WSDM*, 2010.

[15] K. Raynes-Goldie. Pulling sense out of today's informational chaos: Livejournal as a site of knowledge creation and sharing. *First Monday*, 8(12), 2004.

[16] L. Scism and M. Maremont. Insurers test data profiles to identify risky clients. *Wall Street Journal*, Novmber 10 2010.

[17] N. Singer. Face recognition makes the leap from sci-fi. *New York Times*, November 12 2011.

[18] P. Van Mieghem. *Performance Analysis of Communications Networks and Systems*. Cambridge University Press, 2006.

[19] E. Zheleva and L. Getoor. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *18th International World Wide Web Conference*, pages 531–531, April 2009.