Social Link Prediction from Homophilous Relationships Using Probabilistic Logic

Lauc, Davor; Grgic, Sinisa[1]

Abstract

The importance of homophily in network formation is a well-known fact, but link prediction models that use homophilous relations among nodes are rare compared to ones based on the network topology. This paper investigates the possibilities of using available homophilous relations to build a link prediction model using probabilistic logic techniques. A subset of egocentric networks of 10.021 actors was collected, together with basic socio-demographic data from which homophilous relations are deduced. The research demonstrates that the following relations are significant and useful when building a prediction model: physical proximity of actors, having common surname, being coworker, having same place of birth and generational distance. For every relation or combination, the probability function is estimated based on the training subset of the collected dataset. All homophilous relations are combined into an integrative model using a probabilistic logic framework. The predictive power of the model was evaluated and its performance was assessed as excellent.

Keywords:

Social Network Link Prediction, Probabilistic Logics, Homophily, Homophilous Relations Network Formation

1 Introduction

Over the past decades, the importance of homophily in network formation has been well researched. The majority of the existing link prediction models are dominantly based on network topology (Hasan & Zaki, 2011). Research on link prediction models that use homophilous relations among nodes are rare compared to ones based on the network topology, with notable exception of (Hasan, Chaoji, Salem, & Zaki, 2006; Wohlfarth & Ichise, 2008; Patil, Gao, & van de Rijt, 2010; Sachan & Ichise, 2011). Node attributes are often available when network data is incomplete or missing, so it seems worthy to create a prediction model based on homophilous relations and to refine existing network topology models with it. Additionally, node (semantics) similarities have generally lower computational complexity compared to network topology, and have well-developed indexing techniques that enables scalability and makes the model applicable with big data.

People tend to form network ties based on race, age, religion, education, occupation, gender, etc. (Rivera, Soderstrom, & Uzzi, {2010}; Aiello, et al., 2012), according to (McPherson, 2001) following in roughly that order. Other homophilous relations like geographic propinquity, having common surname, company affiliation and other similar relations are also researched and they seem to be a natural context where such ties are formed (Wellman & Berkowitz, 1988; Wellman B. , 1996; Currarini, Jackson, & Pin, 2009).

The main objective of this research was to develop a real-world link prediction model, which can estimate the probability of the link existence between two actors based on their homophilous relations. In this work, homophily is understood in a generalized sense. Any observed equality or similarity between actors (nodes) attributes is considered a homophilous relation.

The link prediction problem can be modeled as a standard supervised classification task and can be tested using existing classification models such as logistic regression, decision trees, naive Bayes or others, where each

[1] Prof.dr.sc. Davor Lauc, Faculty of Social Sciences and Humanities, I.Lucica 3, HR-10000, Zagreb, Croatia; Email addresses: dlauc@ffzg.hr (Lauc, Davor), sinisa.grgic@post.harvard.edu (Grgic, Sinisa)

data point corresponds to a pair of nodes in a social network graph. To improve the real-world applicability of the model, actors attributes that can be mined, extracted, or deduced from publicly available data were favored.

The idea is to use all available homophilous relations (semantic similarities) between nodes that are significant for the existence of a link as predictors, and to build link prediction model using probabilistic logic techniques.

To improve usefulness and applicability of the classification model, one of the most prominent statistical relational learning framework — Problog was used (Raedt, Kimmig, & Toivonen, 2007; Renkens, 2012). This framework enables improvement of a statistical model built by machine learning techniques with expert knowledge and results of previous research, both before and after machine learning process. The probabilistic logic framework also provides easy refinement of the model with other rules and properties, such as network topology. The additional benefit of this framework is that the resulting model is readable increasing its explanatory value.

2 Data

People tend to form many different types of social ties, among whose friendship, kinships and various types of collaboration are the best researched types of links in the area of social network analysis (White, 2011; Wellman & Berkowitz, 1988; Scott & Carrington, 2011; Cingano & Rosolia, 2012). For the purpose of data collection, many operational definitions were used, most of them including sociometric tests (Marsden, 2011). For the purpose of the model developed in this research the operational definition of the link existence based on frequent face-to-face or telephone contact seems to be the most appropriate (Tonnies, 1963). However, face-to-face contact results in many weak ties based on neighborhood, company affiliation and similar homophilous properties, these ties do not have a clear and strong cutoff.

The link existence was operationally defined in order to obtain clear and strong criteria in data collection. It is defined as follows: there should be at least two mobile telecommunication contacts between actors in a given time period (one month). The main advantage of using this requirement is that it is a clear binary empirical notion that can distinguish stronger social ties from accidental ones. Greeting or chatting with your neighbor or workmate does not necessarily result in a strong social tie. Internet social network connections and co-authorships, often used in related researches, do not have the strength of the telecommunication contact. In order to meet this condition, two persons generally need to have exchanged numbers before, answer the phone, and then repeat the call (or repeat to accept the call) at least one more time. Additional practical reasons for such criteria are that participants in the research can answer it without much judgment. There is no requirement from participants to rank different actors or to reason too much about the type/nature of the relationship they were asked for.

The main disadvantage of these criteria in real-world application is that the existence of this type of communications is, although highly correlated, neither necessary nor sufficient condition for "traditional'" social network link types. Because of that, to infer friendship, kinship or some sociometric relationship with certainty, additional data and inference is required.

The main part of data collection was done through on-line surveys, sent to a large number (over 200K) of participants, mainly subscribers to the online services of the supporting partners. Participants were asked to provide general information about themselves (surname, age, current location, education, age of children, hometown, and affiliated company). Then, they were asked to name all people with whom they have had at least two domestic mobile conversations with over the past month. The survey method is similar to a standard name generator, but to ensure data quality and the actors' unique identification possible contacts were limited to the initial list of all participants. About 5% of on-line surveys were filled, after data cleansing there was 10.021 completed surveys - actors in network, and subset of their egocentric network. Among them 2006 network ties (0.2 per actor) were observed. Extrapolated to the whole population, based on initial pool size and a population of 3.5M adult phone users, this indicates a mean of 70 contacts per person per month, which is in the range of telecommunications statistics.

2.1 Limitations of the Dataset

Comparing the demographic of the sample to the whole population, the dataset is slightly biased towards younger participants (36.6 vs 41.7 mean age in 2011 national census) and the male gender (54.3% vs 48.2%). Presumably the bias is due to the method of data collection through on-line surveys, where among Internet users there is a disproportionate number of male and young users, although the exact data are not available. As

the initial participant list is randomly selected from publicly available data and only a small fraction of them completed the on-line survey it is possible that there are other differences between the sample and the population that are not evident. However, due to the relatively large size of the dataset it is expected that these discrepancies do not influence the predictive power of the model too much. The distribution of other parameters that have the available population distribution, like geographical distribution (over 21 counties), are close to the population distribution, with a bootstrap version of the univariate Kolmogorov-Smirnov test p-value of 0.89.

In the context of the social network analysis dataset, which represents only part of the egocentric network and only a tiny fraction of the complete network, standard SNA measurements, such as degree distribution, clustering and various centrality measures, are not applicable. However, for building a predictive classification model, there is large enough collection of positive, connected pairs of actors and negative ones, although with an expected disproportional large number of the latter.

2.1.1 Class skewness

Real networks are generally very sparse. Node degree is usually limited to a small number, and the number of possible links is, in the case of undirected network without loops, $\frac{n(n-1)}{2}$ where n is number of nodes. As number of nodes is large, the probability of link existence between two randomly selected nodes is extremely low. In our sample of $10K$ nodes, with possible number of almost 50 million relationships, and only something more than two thousand links, the probability of link existence is about 2×10^{-5}.

The extreme class skewness of rare events presents problem for standard statistical methods as logistic regression (King & Zeng, 2001) and machine learning classification (Provost, 2000). This is a well-known challenge for link prediction as classification (Backstrom & Leskovec, 2011; Hasan & Zaki, A Survey of Link Prediction in Social Networks, 2011). For the most real-world data, the most accurate, although not very useful prediction, would be to classify all cases as negative, all potential links as non-existing. Therefore, the reasonable objective of the model is not to build binary classifier with high accuracy but to predict the probability of the link existence. Although, estimated probability can be very low, it can be useful in contexts like link recommendation systems and/or it can be combined with network topology.

3 Analysis of Predictors

3.1 Physical proximity of actors

The importance of distance between two actors' residences (neighborhood) is a much debated from the beginnings of social network analysis (Festinger, Back, & Schachter, 1950; Tilahun & Levinson, 2011). For example, (Mok & Wellman, 2007) reported a decrease of the frequency of face-to-face contact if the distance between their addresses exceed five miles, and Wellman at al. (Wellman & Berkowitz, 1988) find that 42% of 'frequent contact' lives within one mile. Many homophilic relationships, including race, social status and others are both cause and effect of physical proximity (Johnston & Pattie, 2011; Bridge, 2002; Ioannides & Topa, 2010). When residence address is known, as often the case of publicly available data, physical proximity can be easily calculated due to widely available GIS data. In this research it is measured as the Euclidean distance between actors who reported their home addresses. Although there are finer measures for neighborhood effect that takes into account configuration of the neighborhood (Doreian, 1981) they are not easily computable from residence addresses alone.

In the dataset, as expected, probability of link existence between two actors is highly negatively correlated with their physical proximity. The probability of defined connection for two persons living on the same address is 10% and it quadraticaly descends with distance (compare Figure 1). The function best fitting this regularity is:

$$p(link|d) = \frac{0.1}{(d+1)^{1.28}}$$

where d is the Eucledian distance in meters. The residual standard error of the function is 0.0001539 on 47,352 degrees of freedom, and the AUC (Area under the ROC Curve) is 0.78.

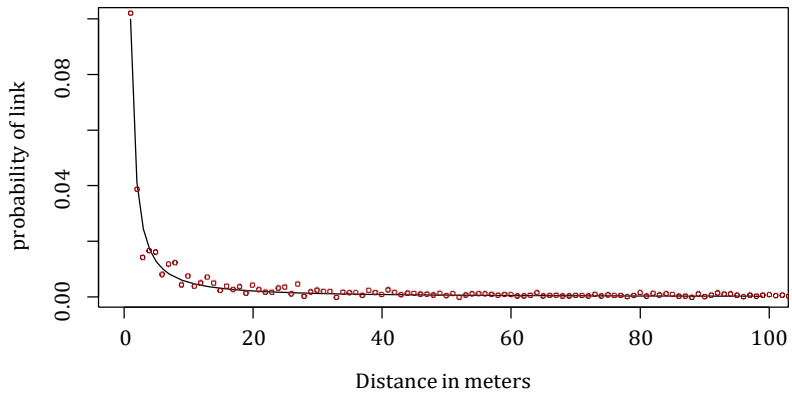

Figure 1: Distance vs. Probability of Link Existence

The relatively low expected value of the link existence for persons sharing an address (d=0) can be explained by the relative rarity of the links in the bigger buildings. Therefore, the obvious additional predictor that moderates relationship between distance and existence of the link is population density. Many researches (Scott & Carrington, 2011), (Sorensen, 2012) demonstrate that such social ties in rural, less populated areas are stronger than in urban areas (compare *gemeinschaft* vs. *gesellschaft* in (Tonnies, 1963). Instead of taking into account the rural – urban continuum (Pahl, 1966) the estimate of the relative population density between actors is used. This enables the model that takes into account the existence of "urban villages" (Beggs, Haines, & Hurlbert, 1996) and similar phenomena.

Relative population density for every pair (a_1, a_2) of actors is estimated as

$$dd = \frac{|\{a_i | d(a_i, a_1) \leq d(a_1, a_2)\}| + |\{a_i | d(a_i, a_2) \leq d(a_1, a_2)\}|}{2}$$

where $d(a_1, a_2)$ is the distance between two actors. Although the minimal dd could be 1 (e.g. two people are the only residents on one address), in the researched sample the minimal dd is 2. In general, to preserve the normalization of the probability, dd should be taken as $max(dd,2)$. The total number of possible actors is approximated by a number of households from a publicly available white page database, and calculated for every actor in the dataset using standard nearest neighbor algorithm. Naturally, population density that is measured in this way is correlated to the distance, with a linear correlation coefficient of 0.789.

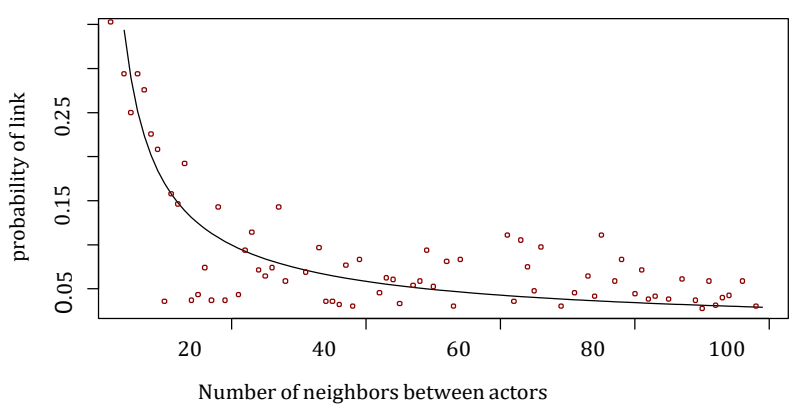

Figure 2: Pop. Density vs. Probability of Link Existence

Although there are some limitations with this measure, as in the case of family houses relatively close to larger buildings, this measure is the better predictor than Euclidean distance. Similarly to distance relationships, there is a negative, polynomial relationship of relative population density to link existence (compare Figure 2). Probability for the relative population density of four (lowest is the dataset) is around 50%. Its decline is approximated with the following:

$$p(link|same\_surname = true, dd) = \frac{1}{d^{0.77}}$$

The fitted function has residual standard error: 0.003233 on 422406 degrees of freedom, and the area under ROC curve is 0.84.

3.2 Common surname

Common surname of two actors is another easily available and possible predictor. In the dataset there are 186 positive examples among 2006 links, resulting in 1.15% probability that two persons having the same last name form a link, versus 0.0036% that they do not. Therefore, a link between two such persons is 315 times more likely, but the probability is too low for any useful prediction.

However, it is naturally expected that this property combined with proximity should predict kinship relations (common households and similar) that is highly correlated with the link existence. Although there are not enough data-points for high quality estimation of the decline of the probability of a link between people sharing the same surname, the following function

$$p(link|dd) = \frac{1}{d^{0.77}}$$

seems reasonable enough (compare Figure **Error! Reference source not found.**). Its residual standard error is 0.1884 on 57 degrees of freedom and AUC is 0.91.

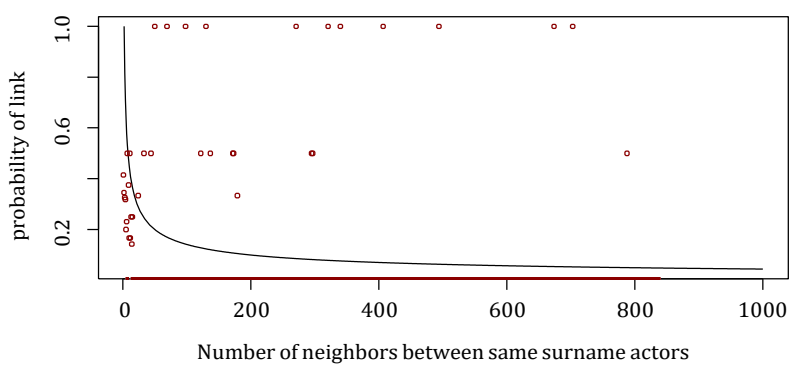

Figure 3: Pop. density of same surnames vs. Link Existence

3.3 Coworkers

Workmates are often compared with neighbors, they are locally available for forming social ties (Wellman B. , 1996). According to Wellman's research workmates make up only 7% of all active community ties, but 26% of these active network members are in contact three times a week or more.

In the dataset there are 442 cases of coworkers forming links, among 104,557 coworkers, resulting in probability of 0.4% that coworkers form link as defined in this research.

Although a higher probability may be expected, there are two natural explanations for it. Firstly, lot of communication related to work is done through face-to-face contact or Internet communication (excluded from our definition). Secondly, strong social ties in larger organizations are rare.

Figure 4: Coworkers by organization size vs. Link Existence

When organization size is taken into account, conditional probability in the dataset is higher for smaller organizations, and is estimated by the function

$$p(link|coworkers = true, num_{cw}) = \frac{0.5}{\sqrt{num_{cw}}}$$

where num_{cw} is the number of employees (coworkers) in the organization (compare Figure 4). This data is taken from the public company registry. The residual standard error of the function is 0.07424 on 970 degrees of freedom. The area under ROC curve is 0.89.

3.4 Same place of birth

Being born in the same city or town, regardless of current residence, is another homophilic relationship that can be categorized under the neighborhood in a wider sense. There is no empirical research supporting this hypothesis, to the authors' knowledge. However, it is self evident that the probability that persons are schoolmates or that they have formed some kind of stronger social ties is higher if they were born in the same place. In the dataset, the probability of link existence for two persons born in the same place is 0.0074%. Obviously, this probability is conditioned upon the size of the hometown. One of the limitations of the place of birth is that people are born in regional hospitals, so that small towns and villages are aggregated. The function that fits data the best is

$$p(link|sametown = true, population_{town}) = \frac{0.02}{\sqrt[4]{population_{town}}}$$

where $population_{town}$ is the population of the place where both actors were born. The residual standard error is 0.001818 on 34 degrees of freedom (the size of the place is rounded to nearest 1000), and the AUC of the function is 0.8289.

3.5 Generational distance

Generational ties are another well known kind of social bonding based on homophilic properties (McPherson, 2001) Hereby calculated as absolute difference in the dates of the birth rounded to the nearest year, the probability of a link in the datasets is presented in the Figure 5.

Overall, generational distance is a significant predictor, starting with probability of about 0.01% for peers born within a six months period and with a quadratic descent of

$$p(link|gd) = \frac{0.001}{\sqrt{gd+1}}$$

where gd represents generational distance starting from 0. The residual standard error of the function is 0.000014 on 57 degrees of freedom.

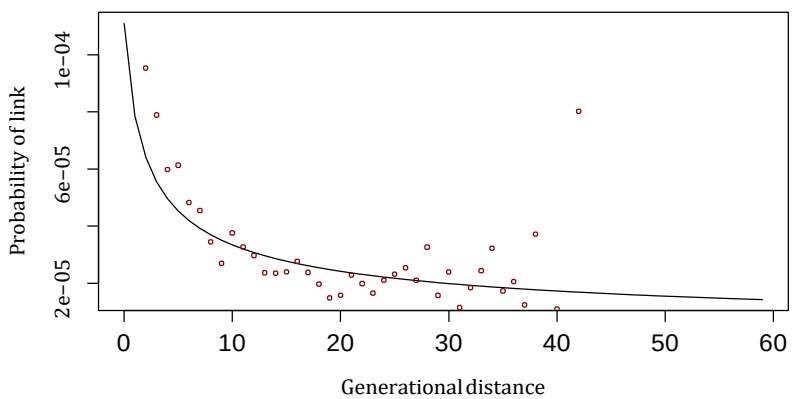

Figure 5: Generational distance

In order to make a more useful prediction geographic propinquity could be used. The probability of ties between peers born within two years distance and who live close together begin with the probability of 25% and descends as

$$p(link|gd \leq 2, dd) = \frac{0.25}{\sqrt{dd-1}}$$

where gd is gen. distance and dd is relative population density as defined above. (compare Figure 6). The residual standard error of the function is 0.00181 on 17642 degrees of freedom, AUC is 0.8458.

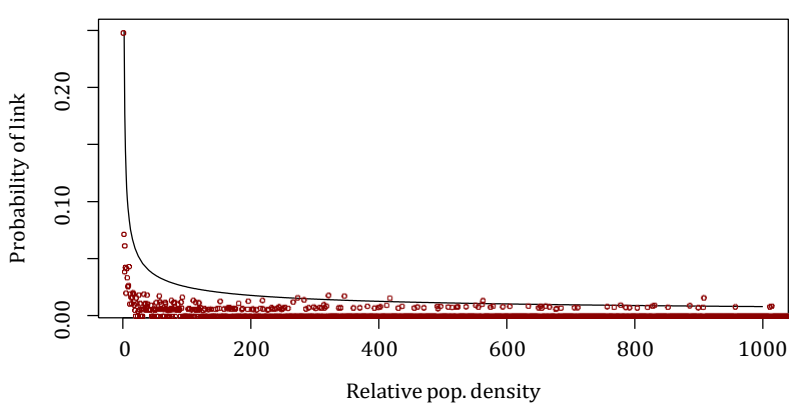

Figure 6: Pop. density where generational distance ≤ 2

4 Combining features in Probabilistic Logic

There are numerous ways to combine the above probabilistic models of link prediction based on separate homophilic relationships into the integrative one, which takes into account all available predictors. One promising approach to link prediction is through using probabilistic logic and statistical relational learning (Popescul & Ungar), (Domingos & Lowd, 2009), (Domingos & Richardson, 2007), (Crane & McDowell, Investigating Markov Logic Networks for Collective Classification., 2012), (Crane & McDowell, Evaluating markov logic networks for collective classification, 2011).

Probabilistic logic can be viewed as a generalization of classical first-order logic. In our context, the classical first-order rule would be, for example,

$$\forall x \forall y (distance(x,y,0) \rightarrow link(x,y))$$

codifying rule that if two persons live at a same address, there is a link between them. However, such rules do not allow exceptions. One pair of unconnected persons living at the same address can falsify the rule, and real-world datasets are, off course, overflowing of such examples. There is hardly any interesting empirical law in social sciences, which allows such hard interpretation.

In probabilistic logic it is possible to attach probability to such rules and soften them, allowing exceptions and regularities that are not law-like. For example, in the dataset there is approximately 0.1 probability of the above rule. So it is much more likely that two persons sharing an address do not form a defined kind of connection, than that they do, however, taking into account the very low general probability of a link, it is still very useful information.

Probabilistic logics have been successfully used in social network analysis and link prediction (Domingos & Richardson, 2007). Their biggest advantage in this context is that they enable not only formalization of researched homophilic probabilistic models, but also relatively easy combination with topological rules when network is (partially) known. For example, rules as transitive closure (links of links are links)

$$\forall x,y,z((link(x,y) \wedge link(y,z)) \rightarrow link(x,z))$$

with a probability depending on how cliquish a network is, are natural extensions of the model. Similar rules can be learned and/or postulated for other link prediction models and techniques [**Error! Reference source not found.**], and other theoretical laws and regularities or "background knowledge constraints" can be easily incorporated into the model.

Another advantage of using probabilistic logics, albeit a more subjective one, is that models in probabilistic logics are “semantically transparent”, at least compared with the state-of-the-art machine learning models such as Support Vector Machines and Neural Networks. Although interaction of the many rules in probabilistic logic can be complicated, and resulting probability distribution not easily readable from the model, there is at least some sense of the regularities and rules that contribute not only to the predictive power but to the explanatory and descriptive power of the model as well.

There are varieties of approaches to probabilistic logic (Getoor & Taskar, 2007), but for the model it seems that two the most appropriate are Markov Logic Networks (Domingos & Richardson, 2007) and Problog (De Raedt, 2008). Although Markov Logic Networks (MLN) support formalization of full first-order logic, and Problog is based on logic programming tradition, due to the scaling problem of MLN, and convenient way to express probabilistic facts with a ‘flexible’ probability of the rules in Problog, the latter was selected.

Rules in Problog are simple Horn clauses with attached probabilities. So, for example the above rule would be expressed as

```
0.1::link(X,Y)  :-  distance(X,Y,0).
```

Problog2 supports flexible intensional probabilities where probability is not prespecified but is an arithmetic expression that needs to be computed. This enables a compact way of formalizing probability distribution of the model’s predictors. For example, rule

```
P::link(X,Y) :-  coworkers(X,Y,N), P is 0.5/(N**0.5).
```

expresses that the probability of link between two persons working together in an organization with N employees is $\frac{0.5}{\sqrt{N}}$

Although all our features are pairwise dependent — e.g. persons born in same town are more likely to work together or live close one to other, there are not enough data-points to estimate joint probability distribution. For practical purposes, it suffices that models should be grouped together based on distance, in particular on relative population density. Those are persons sharing surnames and being of the same generation, so if this is the case their probabilistic model will be used, resulting in the higher probability of links than in the case of two persons having different surnames or not being of the same generation. Therefore, the value of features based on distance, generation and surname will be the maximum of those three.

Interaction with other two predictors is modeled using the noisy-or model,

$$P(y|x_1, \dots, x_n) = 1 - \prod_{i=1}^{n} [\![(1 - P(y|x_i)]\!])$$

which is used by Problog and is implemented in it using following set of the rules:

```
P::link(X,Y) :-  \+sameSurname(X,Y), \+generation(X,Y), dd(X,Y,DD),
                       P is 1.0/(DD**0.77).
P::link(X,Y) :-  sameSurname(X,Y), \+generation(X,Y), dd(X,Y,DD),
                       P is 2**0.5/(DD**0.5).
P::link(X,Y) :-  \+sameSurname(X,Y), generation(X,Y), dd(X,Y,DD),
                       P is 0.25/((DD-1)**0.5).
P::link(X,Y) :-  sameSurname(X,Y), generation(X,Y), dd(X,Y,DD),
                       P is 2**0.5/(DD**0.5).
P::link(X,Y) :-  coworkers(X,Y,N), P is 0.5/(N**0.5).
P::link(X,Y) :-  samehometown(X,Y,N), P is 0.02/(N**0.25).
```

5 Evaluation

The integrative model predicts non-zero probability of the link for about 21% of the all pairs in the test dataset, with a mean value of 0.0001278, however only 0.5% has a probability greater than 1%.

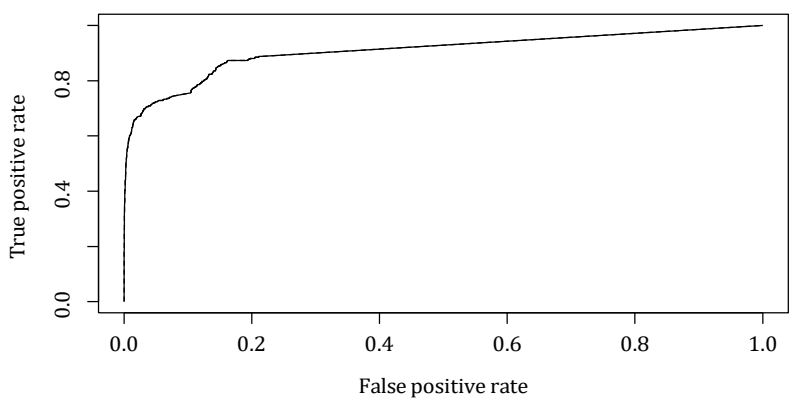

Figure 7: Integrative model - ROC

Due to low general probabilities of the link existence in the social network, the most plausible evaluation criteria of the model seems to be the area under receiver operating characteristic (ROC curve), plotting the fraction of true positive rate vs. false positive rate at various threshold settings. Figure 7 presents ROC curve for the integrative model, for which the area under ROC curve (UAC) is 0.9050, which is excellent (A) performance, evaluated by the traditional academic point system benchmark.

The model is compared with competing models trained on the same data. The support Vector Machine (SVM) performed best among few tested models (GLM, Nearest Neighbor, Naive Bayes, …). Its AUC is 0.8754. Therefore, it may be justified to conclude that using probabilistic logic is right the framework for modeling link prediction from homophilous relationships, even if advantages discussed earlier are not taken into account. Of course, further research using different datasets and in different contexts are needed to support this thesis.

Given low predicted probability and very strict operational definition of link existence, one of the natural questions is how well model generalizes, both to other populations and to other, weaker, definitions of link existence.

6 Conclusions

As for every predictive model, the key theoretical and practical issue is how well the model generalize beyond training data. Given the standard 70%/30% random split into training and testing dataset, it is reasonable to assume that the model is not over-fitted. Other issues are whether the sample is biased to some unknown parameters, and are different populations (non-Internet users, other countries and cultures) significantly different in forming such links. Researches done by (Hofstede, 2001), (Lustig, Koester, & Zhuang, 2006) suggest that there are significant differences among various cultures , and further research is needed to assess and/or generalize the model to different populations.

Other lines of research involve improving the model by increasing the number of homophilous relationships as predictors, where schoolmates, and children of similar age are combined with distance, occupation, and education level, seems to be first choice. Similarly, obtaining more complete samples of the researched network could be used to improve the model with link prediction rules based on network topology.

Another interesting line of research is generalization of the model to types of social links that are different from the one used. It is a safe assumption that if link existence criteria were "do you know person X" or similar, that the number of positive cases among neighbors and workmates would be much larger. Therefore, modeling relation of criteria used in this research to other types of social links could result in interesting theoretical and practical insights.

7 Bibliography

Aggarwal, C. C. (Ed.). (2011). *Social Network Data Analytics.* Springer.

Aiello, L. M., Barrat, A., Schifanella, R., Cattuto, C., Markines, B., & Menczer, F. (2012, #jun#). Friendship prediction and homophily in social media. *ACM Trans. Web, 6*(2), 9:1--9:33. Retrieved from http://doi.acm.org/10.1145/2180861.2180866

Ailshire, J. A. (2010). Do Neighborhood Effects Depend on the Definition of the Neighborhood? *Conference: PAA 2010.*

Al Hasan, M., & Zaki, M. J. (2011). A survey of link prediction in social networks. In *Social network data analytics* (pp. 243-275). Springer.

Backstrom, L., & Leskovec, J. (2011). Supervised random walks: predicting and recommending links in social networks. *Proceedings of the fourth ACM international conference on Web search and data mining* (pp. 635-644). New York, NY, USA: ACM. Retrieved from http://doi.acm.org/10.1145/1935826.1935914

Beggs, J. J., Haines, V. A., & Hurlbert, J. S. (1996). Revisiting the Rural-Urban Contrast: Personal Networks in Nonmetropolitan and Metropolitan Settings1. *Rural sociology, 61*(2), 306-325.

Borgatti, S. P., & Everett, M. G. (2006). A graph-theoretic perspective on centrality. *Social networks, 28*(4), 466-484.

Bridge, G. (2002). The neighbourhood and social networks. *Bristol: ESR Centre*.

Chawla, N. V. (2002). Smote: Synthetic minority over-sampling technique. *Journal of A.I. Research, 16*, 341–378.

Cingano, F., & Rosolia, A. (2012). People I know: job search and social networks. *Journal of Labor Economics, 30*(2), 291-332.

Crane, R., & McDowell, L. (2012). Investigating Markov Logic Networks for Collective Classification. *ICAART (1)*, (pp. 5-15).

Crane, R., & McDowell, L. K. (2011). Evaluating markov logic networks for collective classification. *Proceedings of the 9th MLG Workshop at the 17th ACM SIGKDD Conference on Knowledge Discovery and Data Mining.*

Currarini, S., Jackson, M. O., & Pin, P. (2009). An economic model of friendship: Homophily, minorities, and segregation. *Econometrica, 77*(4), 1003-1045.

D. Fierens, G. V., & Raedt., L. D. (2013). Inference and learning in probabilistic logic programs using weighted Boolean formulas. *Theory and Practice of Logic Programming, arXiv preprint arXiv:1304.6810 (to appear in),.*

De Raedt, L. (2008). *Logical and Relational Learning.* Springer. Retrieved from http://books.google.com/books?id=0EYVINC1paAC

Domingos, P., & Lowd, D. (2009). *Markov Logic: An Interface Layer for Artificial Intelligence.* Morgan \& Claypool. Retrieved from http://books.google.com/books?id=ijqFfoIy_T0C

Domingos, P., & Richardson, M. (2007). 1 Markov Logic: A Unifying Framework for Statistical Relational Learning. *STATISTICAL RELATIONAL LEARNING*, 339.

Doreian, P. (1981). Estimating linear models with spatially distributed data. *Sociological methodology, 12*, 359-388.

Festinger, L., Back, K., & Schachter, S. (1950). *Social Pressures in Informal Groups: A Study of Human Factors in Housing.* Stanford University Press. Retrieved from http://books.google.com/books?id=1zSsAAAAIAAJ

Getoor, L., & Taskar, B. (2007). *Introduction to Statistical Relational Learning.* MIT Press. Retrieved from http://books.google.com/books?id=lSkIewOw2WoC

Hasan, M. A., & Zaki, M. J. (2011). A Survey of Link Prediction in Social Networks. In *Social Network Data Analytics* (pp. 243-275).

Hasan, M. A., Chaoji, V., Salem, S., & Zaki, M. (2006). Link prediction using supervised learning. *In Proc. of SDM 06 workshop on Link Analysis, Counterterrorism and Security.*

Hofstede, G. H. (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations.* Sage.

Ioannides, Y. M., & Topa, G. (2010). NEIGHBORHOOD EFFECTS: ACCOMPLISHMENTS AND LOOKING BEYOND THEM*. *Journal of Regional Science, 50*(1), 343-362.

Johnston, R., & Pattie, C. (2011). Social networks, geography and neighbourhood effects. In J. Scott, & P. Carrington (Eds.), *The SAGE Handbook of Social Network Analysis* (pp. 301-311). SAGE Publications Ltd.

King, G., & Zeng, L. (2001). Logistic regression in rare events data. *Political analysis, 9*(2), 137-163.

Lazarsfeld, P., as a Social Process: A Substantive, R. M., & Control in Modern Societv M. Berger, e. N. (1954). Friendship as a Social Process: A Substantive. In M. Berger (Ed.), *Freedom and Control in Modern Society.* New York: Van Nostrand.

Lowd, D., & Domingos, P. (2007). Efficient weight learning for Markov logic networks. *In Proceedings of the Eleventh European Conference on Principles and Practice of Knowledge Discovery in Databases*, (str. 200-211).

Lustig, M. W., Koester, J., & Zhuang, E. (2006). *Intercultural competence: Interpersonal communication across cultures.* Pearson/A and B.

Marsden, P. V. (2011). Survey Methods for Network Data. In J. Scott, & P. J. Carrington (Eds.), *The Sage Handbook of social Network Analysis* (pp. 370-388). London: Sage Publications.

McPherson, M. e. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology, 27*, 415-444.

Mok, D., & Wellman, B. (2007). Did distance matter before the Internet?: Interpersonal contact and support in the 1970s. *Social networks, 29*(3), 430-461.

Pahl, R. E. (1966). THE RURAL-URBAN CONTINUUM1. *Sociologia ruralis, 6*(3), 299-329.

Park, R. E., & Wirth, L. (1925). *The City.* Chicago, IL: University of Chicago Press.

Patil, A., Gao, J., & van de Rijt, A. (2010). Homophily Based Link Prediction in Friendship Networks. *The International Sunbelt Social Networks Conference XXX, Trento, Italy, June 29-July 4 2010.*

Popescul, A., & Ungar, L. H. (n.d.). Statistical relational learning for link prediction.

Provost, F. (2000). Machine learning from imbalanced data sets 101. *Proceedings of the AAAI'2000 workshop on imbalanced data sets.*

Raedt, L. D., Kimmig, A., & Toivonen, H. (2007). ProbLog: a probabilistic Prolog and its application in link discovery. *In Proceedings of 20th International Joint Conference on Artificial Intelligence* (pp. 2468-2473). AAAI Press.

Renkens, J. S. (2012). ProbLog2: From probabilistic programming to statistical relational learning,. *Proceedings of the NIPS Probabilistic Programming Workshop, Lake Tahoe, USA, 7-8 December 2012.*

Richardson, M., & Domingos, P. (#feb# 2006). Markov logic networks. *Mach. Learn., 62*(1-2), 107-136. Dohvaćeno iz http://dx.doi.org/10.1007/s10994-006-5833-1

Rivera, M. T., Soderstrom, S. B., & Uzzi, B. ({2010}). {Dynamics of Dyads in Social Networks: Assortative, Relational, and Proximity Mechanisms}. In K. {Cook, & D. Massey (Eds.), *{ANNUAL REVIEW OF SOCIOLOGY, VOL 36}* (Vol. {36}, pp. {91-115}). {4139 EL CAMINO WAY, PO BOX 10139, PALO ALTO, CA 94303-0897 USA}: {ANNUAL REVIEWS}.

Sachan, M., & Ichise, R. (2011). Using semantic information to improve link prediction results in network datasets. *International Journal of Computer Theory and Engeneering, 3*, 71-76.

Scott, J., & Carrington, P. J. (2011). *The SAGE handbook of social network analysis.* SAGE publications.

Sorensen, J. F. (2012). Testing the hypothesis of higher social capital in rural areas: the case of Denmark. *Regional Studies, 46*(7), 873-891.

Tilahun, N., & Levinson, D. (2011). Work and home location: Possible role of social networks. *Transportation Research Part A: Policy and Practice, 45*(4), 323-331.

Tonnies, F. (1963). Community and Society (1887). *NY: Harper \& Row*.

Wellman, B. (1996). Are personal communities local? A Dumptarian reconsideration. *Social networks, 18*(4), 347-354.

Wellman, B. S., & Berkowitz, S. D. (1988). *Social structures: A network approach* (Vol. 2). CUP Archive.

White, D. R. (2011). Kinship, Class and Community. *JC Scott \& P. Carrington ({\'e}ds.) Sage Handbook of Social Networks. New York, Sage Publications*, 129-147.

Wohlfarth, T., & Ichise, R. (2008). Semantic and event-based approach for link prediction. In *Practical Aspects of Knowledge Management* (pp. 50-61). Springer.