Supervised Link Prediction by Social and Demographic Attributes

Jeongjin Ku
jeongjin@stanford.edu

Wonhong Lee
wonhong@stanford.edu

Department of Computer Science
Stanford University, CA 94305

Abstract

In this paper, we employ various learning algorithms to develop an efficient link prediction model based on social and demographic attributes.

Contents

1 Introduction

Link prediction in complex network is an active area of research in network analysis. This task is complicated by the fact that shape dynamics of the network is constantly changing, and it is difficult to define which inherent factors drive this change. In this paper, we will complement an existing algorithm by considering social, geographic, and demographic features to enhance the outcome of link predictions.

1.1 Related Work

Backstrom and Leskovec [4] devised a link prediction model based on features involving personal attributes and network topology. This algorithm, however, may be infeasible in many cases due to limitations in acquiring personal attributes. With a heightened awareness towards privacy issues, it has become more difficult to collect personal information.

The model introduced by Liben-Nowell and Keinberg [3] predicts possible connections between nodes in a social network based on graph theoretic measures. Although this algorithm is effective in making predictions on existing nodes, we do not have the same assurance for newly formed nodes since they do not hold any network topological information.

The main idea in the prediction model proposed by Scellato [5] is to incorporate geographic features such as physical distance and check-in data. This is a promising approach which demonstrates how features other than network topology and social attributes can be relevant in link prediction. We believe that there are new ways to render such geographic information to improve prediction results.

1.2 Problem Formulation

We want to develop a robust algorithm that can make accurate predictions for both existing and newly formed nodes. Some graph theoretic measures, such as the Adamic-Adar score, play a crucial role in link prediction, and they often yield outstanding results for existing nodes. We will certainly incorporate these features in building a new prediction model.

As mentioned above, however, it is difficult to make predictions for newly formed nodes by solely analyzing the graph theoretic properties. We run into similar obstacles when predicting possible links between a pair of nodes with distance greater than 2. It is easy to see how conventional notions of

topology might be insufficient for a meaningful prediction. To address such limitations, we complement topological features by employing new features of social, geographic, and demographic flavor.

2 Data Rendering

In this paper, we consider two types of data sets, namely social network and geographic data. We obtained the first data set from Gowalla, an online location-based social network owned by Facebook. The second data set is the 2000 United States Census, which consists of demographics for each ZIP code area.

2.1 Social Network Data

For the first data set, we have friendship snapshots taken at July of 2010 and October of 2010, and public check-in history on February of 2009 and October of 2010 for users worldwide.

Not only is the size of the first data set massive, but the demographic information from the second data set is restricted to the United States. We therefore extract information on the set of users with at least one check-in point in the United States. The following table shows the size of the reduced data set.

Time of snapshot	Number of nodes	Number of edges
July of 2010	26,989	115,495
October of 2010	36,231	178,791

As for the check-in history, there are 3,742,003 locations for the two time frames combined.

For simplicity, let t_1 and t_2 denote the time, in chronological order, at which friendship snapshots were taken. We will also refer to users as nodes. Each node is uniquely assigned to a nonnegative integer.

Since the adjacency matrix for the reduced data set is extremely sparse, we only consider the set $\mathcal{A}$ of nodes incident to an edge that is present at t_2 but not at t_1. In other words, $\mathcal{A}$ consists of active nodes. We further define $\mathcal{A}_1$ as the subset of $\mathcal{A}$ with nodes present only at t_1, and set $\mathcal{A}_2 = \mathcal{A} \setminus \mathcal{A}_1$. Hence, $\mathcal{A}_2$ is the set of newly formed nodes. Note also that $\mathcal{A}_1$ and $\mathcal{A}_2$ are disjoint subsets of $\mathcal{A}$.

Let u, v, and w be nodes in $\mathcal{A}$. The degree of u is denoted $\deg u$. We write $w \sim \{u, v\}$ when w is adjacent to both u and v. The distance between u and v is written as $d(u, v)$.

The set of all check-in locations for both time frames will be denoted by Λ. Hence, we do not distinguish between check-in points visited in different time frames. We also define $\lambda(u)$ to be the set of all check-in locations of $u \in \mathcal{A}$. Note that Λ is the disjoint union of $\lambda(u)$ for all $u \in \mathcal{A}$.

Furthermore, we select subsets of $\mathcal{A} \times \mathcal{A}$ from which we plan to build the training examples.

$$\begin{aligned}
\mathcal{S}_2 &= \{(u, v) \in \mathcal{A}_1 \times \mathcal{A}_1 : d(u, v) = 2 \text{ with } u > v\} \\
\mathcal{S}_3 &= \{(u, v) \in \mathcal{A}_1 \times \mathcal{A}_1 : d(u, v) = 3 \text{ with } u > v\} \\
\mathcal{S}_\infty &= \{(u, v) \in \mathcal{A}_1 \times \mathcal{A}_2 : u \sim v\}.
\end{aligned}$$

We use $\mathcal{S}_2$ to consider the case when we can use topological, geometric, and social features. As for $\mathcal{S}_3$, we want to investigate the performance of the prediction model when we cannot make use of the information provided by the network topology. Finally, by examining the case of $\mathcal{X}_\infty$, we will be able to test link predictions for newly formed nodes. Hence, by taking into account, various training examples, we hope to demonstrate that the prediction model performs well under certain limitations.

2.2 Demographic Data

The second data set consists of 19 fields for various demographical attributes. The following table lists some of these features relevant to link prediction.

ZIP code	Area code
Population	Total population
Population density	Population per unit area
Geographic area	Urban, suburban, farm, non-farm
Race	White, black, Asian, Indian, Hawaiian, other
Age	Age groups
Education	Education level of population over 18
Household income	Median household income
Per capita income	Median income per person
House value	Average value of homes
Housing density	Number of houses per unit area

For consistency, the check-in locations in the first data set are converted into an area code by using the Geo-postal Service provided by Nuestar.

3 Feature Selection

The features we define can be classified into three categories depending on whether they are relevant to topological, geographic, or social attributes. We then carry out feature selection by computing the Kullback-Leibler divergence of each feature.

3.1 Topological Features

The topological features are by far the most important features as they retain information on the graph theoretical properties of the network. The most natural topological feature is the number of common nodes, denoted τ_n. That is, given $u \in \mathcal{A}$ and $v \in \mathcal{A}$,

$$\tau_n(u, v) = \sum_{w \notin \{u,v\}} \mathbf{1}(w \sim \{u, v\}).$$

Observe that this feature does not take into account the fact that users corresponding to nodes with higher degree are more likely to be friends with a larger group of users.

The cosine similarity τ_c of $u, v \in \mathcal{A}$ is defined as

$$\tau_c(u, v) = \frac{\tau_n(u, v)}{\deg u \cdot \deg v}.$$

By incorporating this feature into the prediction model, we lend less significance to a pair of nodes with higher degree since users corresponding to these two nodes are more likely to have common friends.

The Adamic-Adar score τ_a of $u, v \in \mathcal{A}$ is given by

$$\tau_a(u, v) = \sum_{w \sim \{u,v\}} \frac{1}{\log(\deg w)}.$$

We employ this feature to downgrade the effect of common nodes with higher degree since users corresponding to these nodes are more likely to be friends with a larger group of users.

We also define the preferential attachment τ_p of $u \in \mathcal{A}$ and $v \in \mathcal{A}$ as

$$\tau_p(u, v) = \deg u \cdot \deg v.$$

This feature captures more active users corresponding to nodes with higher degree.

3.2 Geograhpic Features

We now define a set of geographic features based on the check-in history. Each check-in point is a physical location which can be written in the geographic coordinate system, that is, for $x \in \lambda(u)$ for some $u \in \mathcal{A}$,

$$\gamma_p(x) = (\theta, \phi),$$

where θ and ϕ are the latitude and longditude of x, respectively.

The mode γ_m of $u \in \mathcal{A}$ is given by

$$\gamma_m(u) = \arg\max_{x \in \lambda(u)} \mathbf{P}(x(u)),$$

that is, the check-in location of u that occurs most frequently.

Similarly, the sample mean γ_s of $u \in \mathcal{A}$ is defined in the usual way as

$$\gamma_s(u) = \frac{\sum_{x \in \lambda(u)} x}{\sum_{x \in \Lambda} \mathbf{1}(x \in \lambda(u))},$$

that is, the arithmetic mean of the check-in locations of u.

We would also like to define a feature that captures the intuition of communities within a network. To do this, we repeatedly apply k-means clustering to form a binary decision tree for each $\lambda(u)$. Among the leaf clusters from the decision tree, we choose the cluster with the most check-in locations, and denote its mean as $\mu(u)$. Then the clustering distance between $u \in \mathcal{A}$ and $v \in \mathcal{A}$ is defined as

$$\gamma_c(u, v) = \|\mu(u) - \mu(v)\|_2,$$

that is, the Euclidean distance between the mean of the largest leaf clusters in each decision tree.

3.3 Social Features

Although 27 social features are considered in the prediction model, we only discuss a few of the important ones as the rest are defined similarly. We write $N(u)$ to denote the total population of the area code for $u \in \mathcal{A}$.

We define the housing density σ_h of $u \in \mathcal{A}$ as

$$\sigma_h(u) = \frac{H(u)}{A(u)},$$

where $H(u)$ is the number of houses in the area code for u.

The density of white population σ_w for $u \in \mathcal{A}$ is given by

$$\sigma_w(u) = \frac{W(u)}{N(u)},$$

where $W(u)$ is the white population of the area code for u.

We write the per capita income σ_p of $u \in \mathcal{A}$ as

$$\sigma_p(u) = \frac{I(u)}{N(u)},$$

where $I(u)$ is the net income of residents in the area code for u.

Finally, the urban population density σ_u of $u \in \mathcal{A}$ is defined as

$$\sigma_u(u) = \frac{U(u)}{N(u)},$$

where $U(u)$ is the population living in urban areas within the area code for u.

3.4 Chi-square Score

We compute the chi-square score for each feature, and then, enumerate these features in descending order.

Feature	χ^2 Score	Feature	χ^2 Score	Feature	χ^2 Score
τ_n	549.07	γ_p	1,995.87	σ_h	680.34
τ_c	614.85	γ_m	4,076.83	σ_w	659.14
τ_a	2,484.30	γ_s	2,965.92	σ_p	522.94
τ_p	4,429.33	γ_c	4,210.23	σ_u	661.89

According to this table, the preferential attachment τ_p has the highest chi-square score. This may be due to the fact that Gowalla was at its early stage at the time the data was collected.

Another interesting observation to make is that the mode γ_m and the clustering distance γ_c are ranked second and third, respectively. This appears to reflect the fact that geographical features for newly formed nodes are highly correlated to their respective labels with respect to its adjacent nodes. We also observe that the social feature have significant chi-squre scores, which indicate that they will enhance the prediction results.

Scellato[5] uses a prediction model based on geographic features such as the mode and mean of check-in locations. We devised a new distance metric, namely γ_c, by using decision-tree clustering. As we can see in the χ^2 score table, clustering distance feature shows better performance than γ_m and γ_s.

4 Supervised Learning Results

In this section, we build training sets $\mathcal{X}_2$, $\mathcal{X}_3$, and $\mathcal{X}_\infty$ based on $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_\infty$, respectively. By then apply learning algorithms on these traing sets, and make link predictions. By considering these three cases, we hope to demonstrate that this model is able to make meaningful predictions for newly formed nodes when graph theoretic features are not available. We will also verify that the geographical and social features play an important role when network topological information is not available.

4.1 Training Examples

We defined $\mathcal{S}_2$, $\mathcal{S}_3$, and $\mathcal{S}_\infty$ in Section 2. Now let τ, γ, and σ, resepctively, denote the set of topological, geographical, and social features discussed in Section 3. Then we define the training sets $\mathcal{X}_2$, $\mathcal{X}_3$, and $\mathcal{X}_\infty$ as

$$\mathcal{X}_2 = \mathcal{S}_2 \cup \tau \cup \gamma \cup \sigma$$
$$\mathcal{X}_3 = \mathcal{S}_3 \cup \{\tau_p\} \cup \gamma \cup \sigma$$
$$\mathcal{X}_\infty = \mathcal{S}_\infty \cup \{\tau_p\} \cup \gamma \cup \sigma$$

We note that $\mathcal{X}_3$ and $\mathcal{X}_\infty$ are independent of τ, the set of graph theoretic features.

Since $d(u, v) = 2$ for $u, v \in \mathcal{S}_2$, we can make use of features in τ when learning $\mathcal{X}_2$. In contrast, we cannot take advantage of features in τ when learning $\mathcal{X}_3$ and $\mathcal{X}_\infty$ except for τ_p. This is because, for $u, v \in \mathcal{S}_3$, we have that $d(u, v) = 3$, and so, we do not have any common neighbors for u and v. As for $\mathcal{S}_\infty$, we are dealing with newly formed nodes, and so, we cannot use features in τ, which are based on topological information at t_1.

Hence, by considering the three training examples, we want to see if the prediction model can overcome the following obstacles.

$\mathcal{X}_2$	When all features τ, γ, σ are available
$\mathcal{X}_3$	When we cannot use $\tau \backslash \{\tau_p\}$
$\mathcal{X}_\infty$	When users are newly added in the social network

We now investigate these three cases separately.

4.2 Prediction Results on $\mathcal{X}_2$

Before we run learning algorithms on the entire training set $\mathcal{X}_2$, we want to observe how addition of geographical and social features enhances prediction

results. Hence, we will consider the following subsets of the trainset $\mathcal{X}_2$.

$$\mathcal{X}_2'' = \mathcal{S}_2 \cup \tau \text{ and } \mathcal{X}_2' = \mathcal{S}_2 \cup \tau \cup \gamma.$$

We will learn in the order of $\mathcal{X}_2''$, $\mathcal{X}_2'$, and $\mathcal{X}_2$, and see how adding a new set of features improves the performance of the overal prediction algorithm.

Learning set	Random forest	Decision tree J48	Adaboost M1
$\mathcal{X}_2''$	70.60%	71.80%	66.80%
$\mathcal{X}_2'$	77.50%	78.00%	73.60%
$\mathcal{X}_2$	80.20%	78.60%	75.90%

(F1 score)

The results in the table shows that the addition of τ and γ into the training set improves prediction performance for all machine-learning classifiers. For the random forest classifier, the F1 score increases from 70.6% on $\mathcal{X}_2''$ to 77.5% on $\mathcal{X}_2'$, which is a 6.9% improvement. This demonstrates that geographical features play an important role when making link predictions. Finally, for $\mathcal{X}_2$, the F1 score is 80.2%, which is 10% and 2.5% greater than the case of $\mathcal{X}_2''$ and $\mathcal{X}_2'$, respectively.

4.3 Prediction Results on $\mathcal{X}_3$

As in the previous section, we gradually increase the size of the training set. We will consider the following subset of $\mathcal{X}_3$.

$$\mathcal{X}_3'' = \mathcal{S}_3 \cup \{\tau_p\} \text{ and } \mathcal{X}_3' = \mathcal{S}_3 \cup \{\tau_p\} \cup \gamma.$$

Hence, we run learning algorithms on $\mathcal{X}_3''$, $\mathcal{X}_3'$, and $\mathcal{X}_3$, in chronological order. Note that, since we are considering pairs of nodes of distance 3, we do not incorporate topological features such as the Adamic-Adar score and cosine similarity. We only consider $\tau_p \in \tau$.

Learning set	Random forest	Decision tree J48	Adaboost M1
$\mathcal{X}_3''$	65.30%	56.60%	57.60%
$\mathcal{X}_3'$	70.90%	71.70%	64.80%
$\mathcal{X}_3$	71.30%	70.00%	64.90%

(F1 score)

When running random forest classifier, there is a improvement of approximately 5% in the F1 score when training on $\mathcal{X}_3'$ instead of $\mathcal{X}_3''$, that is, by

addition of geographic features γ. Furthurmore, the F1 score on $\mathcal{X}_3$ is 71.3%, and so, by adding the social features σ, the F1 score increases by 6%. This is a remarkable result since we are hardly relying on topological features to train the prediction algorithm. We see that in the absence of topological features, geographic and social features play an important role in enhancing the prediction results. We note that, since we cannot make use of important features of τ such as the Adamic-Adar score, we see that the overall result for $\mathcal{X}_3$ is slightly lower than that of $\mathcal{X}_2$.

4.4 Prediction Results on $\mathcal{X}_\infty$

We now investigate the results for $\mathcal{X}_\infty$ by first running the learning algorithms on the following subsets.

$$\mathcal{X}''_\infty = \mathcal{S}_\infty \cup \{\tau_p\} \text{ and } \mathcal{X}'_\infty = \mathcal{S}_\infty \cup \{\tau_p\} \cup \gamma.$$

We will consider $\mathcal{X}''_\infty$, $\mathcal{X}'_\infty$ and $\mathcal{X}_\infty$ in chronological order. This is a particularly important case since we are testing the performance of the prediction model on newly formed nodes.

Learning set	Random forest	Decision tree J48	Adaboost M1
$\mathcal{X}''_\infty$	47.20%	31.00%	31.10%
$\mathcal{X}'_\infty$	66.40%	63.10%	65.00%
$\mathcal{X}_\infty$	82.80%	79.90%	81.20%

(F1 score)

As with the first two cases, we see an increase in F1 score for all three learning classifiers. For the random forest classifer in specific, we see that the F1 scores increases from 47.20% on $\mathcal{X}''_\infty$ to 66.40% on $\mathcal{X}'_\infty$, which is a 19.20% increase. We have an F1 score of 82.80% on $\mathcal{X}_2$ which is a 16.40% increase. Surprisingly, the results on $\mathcal{X}_\infty$ are generally better than that of $\mathcal{X}_2$ and $\mathcal{X}_3$. This abnormality may be explained by the fact that the friendship snapshots were taken soon after the launch of Gowalla. Nonetheless, this demonstrates that geographic, demographic, and social features play a critical role in improving link prediction performance.

4.5 Conclusion

In defining geographic features, we defined a new feature, namely γ_c, which is the clustering mean based on the decision tree. By computing the chi-square

scores of the features, we were able to verify that γ_c had higher scores than other geometric features. We also devised a new way of incorporating social features by making use of the United States Census data.

As we can see from the table below, the prediction model performs better on $\mathcal{X}_2''$ compared to $\mathcal{X}_3''$ and $\mathcal{X}_\infty''$. However, we see that the F1 scores for $\mathcal{X}_2$, $\mathcal{X}_3$, and $\mathcal{X}_\infty$ are comparable. Note also that, by comparing F1 scores on $\mathcal{X}_\infty''$ and $\mathcal{X}_\infty$, we see that geographic, demographic, and social features are a powerful tool in making link predictions for newly formed nodes. This shows that geographic, demographic, and social features can make up for the shortcoming of network topological features.

Learning set	$\mathcal{X}_2$	$\mathcal{X}_3$	$\mathcal{X}_\infty$
$\mathcal{X}''$	70.60%	65.30%	47.20%
$\mathcal{X}$	80.20%	71.30%	82.80%

(F1 score on random forest classifier)

Observe that there is approximately a 10% improvement from $\mathcal{X}_2''$ to $\mathcal{X}_2$. This implies that geographic and social features plays an important role even when graph theoretic information is not available. Heuristically, it appears that γ and σ are nearly "orthogonal" to τ in the sense that the addition of γ and σ into the training example greatly increases the F1 score. We can infer from this that γ and σ provide meaningful information vastly different from that provided by τ.

References

[1] Arvind Narayanan et al *De-anonymizing Social Networks*. 2009 .

[2] Akshay N Patil, *Homophily based link prediction in social networks*. 2004.

[3] David Liben-Nowell et al, *The Link Prediction Problem for Social Networks*. 2004.

[4] Lars Backstrom et al *Supervised Random Walks: Predicting and Recommending Links in Social Networks* 2011.

[5] Salvatore Scellato et al, *Exploiting place features in link prediction on location-based social networks*. 2011.

[6] David Liben-Nowell et al, *Geographic routing in social networks*. 2005.