

Joint Link Prediction and Attribute Inference Using a Social-Attribute Network

NEIL ZHENQIANG GONG, AMEET TALWALKAR, and LESTER MACKEY,

University of California, Berkeley

LING HUANG, Intel Labs

EUI CHUL RICHARD SHIN and EMIL STEFANOV, University of California, Berkeley

ELAINE (RUNTING) SHI, University of Maryland, College Park

DAWN SONG, University of California, Berkeley

The effects of social influence and homophily suggest that both network structure and node-attribute information should inform the tasks of link prediction and node-attribute inference. Recently, Yin et al. [2010a, 2010b] proposed an attribute-augmented social network model, which we call *Social-Attribute Network* (SAN), to integrate network structure and node attributes to perform both link prediction and attribute inference. They focused on generalizing the random walk with a restart algorithm to the SAN framework and showed improved performance. In this article, we extend the SAN framework with several leading supervised and unsupervised link-prediction algorithms and demonstrate performance improvement for each algorithm on both link prediction and attribute inference. Moreover, we make the novel observation that attribute inference can help inform link prediction, that is, link-prediction accuracy is further improved by first inferring missing attributes. We comprehensively evaluate these algorithms and compare them with other existing algorithms using a novel, large-scale Google+ dataset, which we make publicly available (<http://www.cs.berkeley.edu/~stevgong/gplus.html>).

Categories and Subject Descriptors: K.4.1 [Computers and Society]: Public Policy Issues—*Privacy*; H.2.8 [Database Management]: Database Applications—*Data mining*

General Terms: Algorithms, Measurement

Additional Key Words and Phrases: Link prediction, attribute inference, social-attribute network, heterogeneous network, Google+

ACM Reference Format:

Neil Zhenqiang Gong, Ameet Talwalkar, Lester Mackey, Ling Huang, Eui Chul Richard Shin, Emil Stefanov, Elaine (Runting) Shi, and Dawn Song. 2014. Joint link prediction and attribute inference using a social-attribute network. *ACM Trans. Intell. Syst. Technol.* 5, 2, Article 27 (April 2014), 20 pages.
DOI: <http://dx.doi.org/10.1145/2594455>

This work is supported by the NSF under grants CCF-0424422, 0311808, 0832943, 0448452, 0842694, 0627511, 0842695, 0808617, 1122732, and 0831501 CT-L, by the AFOSR under MURI award no. FA9550-09-1-0539, by the AFRL under grant no. P010071555, by the Office of Naval Research under MURI grant no. N000140911081, by the MURI program under AFOSR grant no. FA9550-08-1-0352, the NSF Graduate Research Fellowship under grant no. DGE-0946797, by the DoD through the NDSEG Program, by Intel through the ISTC for Secure Computing, and by a grant from the Amazon Web Services in Education program. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Authors' addresses: N. Z. Gong (corresponding author), A. Talwalkar, L. Mackey, R. Shin, E. Stefanov, and D. Song, Computer Science Division, Soda Hall, University of California at Berkeley, Berkeley, CA 94720; L. Huang, Intel Labs, ISTC on Secure Computing at UC Berkeley, 711 Soda Hall, Berkeley, CA 94720. E. Shi, Department of Computer Science, University of Maryland, College Park; corresponding author's email: neilz.gong@berkeley.edu.

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.

© 2014 ACM 2157-6904/2014/04-ART27 \$15.00

DOI: <http://dx.doi.org/10.1145/2594455>

1. INTRODUCTION

Online social networks (e.g., Facebook, Google+) have become increasingly important resources for interacting with people, processing information, and diffusing social influence. Understanding and modeling the mechanisms by which these networks evolve are therefore fundamental issues and active areas of research.

The classical *link-prediction problem* [Liben-Nowell and Kleinberg 2003; Gong et al. 2012a] has attracted particular interest. In this setting, we are given a snapshot of a social network at time t and aim to predict links (e.g., friendships) that will emerge in the network between t and a later time t' . Alternatively, we can imagine the setting in which some links existed at time t but are missing at t' . In online social networks, a change in privacy settings often leads to missing links, for example, a user on Google+ might decide to hide her family circle between times t and t' . The missing-link problem has important ramifications, as missing links can alter estimates of network-level statistics [Kossinets 2006], and the ability to infer these missing links raises serious privacy concerns for social networks. Since the same algorithms can be used to predict new links and missing links, we refer to these problems jointly as *link prediction*.

Another problem of increasing interest revolves around node attributes [Zheleva and Getoor 2009; Gong et al. 2012a]. Many real-world networks contain rich categorical node attributes, for example, users in Google+ have profiles with attributes including employer, school, occupation, and places lived. In the *attribute-inference problem*, we aim to populate attribute information for network nodes with missing or incomplete attribute data. This scenario often arises in practice when users in online social networks set their profiles to be publicly invisible or create an account without providing any attribute information. The growing interest in this problem is highlighted by the privacy implications associated with attribute inference as well as the importance of attribute information for applications, including people search,¹ collaborative filtering [Melville and Sindhvani 2010], and user identity resolution [Bartunov et al. 2012].

In this work, we simultaneously use network structure and node-attribute information to improve performance of both the link-prediction and the attribute-inference problems, motivated by the observed interaction and homophily between network structure and node attributes. The principle of social influence [Fond and Neville 2011], which states that users who are linked are likely to adopt similar attributes, suggests that network structure should inform attribute inference. Other evidence of interaction [Kumar et al. 2004; Kim and Leskovec 2011] shows that users with similar attributes, or in some cases antithetical attributes, are likely to link to one another, motivating the use of attribute information for link prediction. Additionally, previous studies [Kossinets and Watts 2006; Fond and Neville 2011] have empirically demonstrated those effects on real-world social networks, providing further support for considering both network structure and node attribute information when predicting links or inferring attributes.

However, the algorithmic question of how to simultaneously incorporate these two sources of information remains largely unanswered. Relational learning [Taskar et al. 2003; Miller et al. 2009; Yu et al. 2006; Singh and Gordon 2008], matrix factorization, and alignment [Menon and Elkan 2011; Scripps et al. 2009] based approaches, have been proposed to leverage attribute information for link prediction, but they suffer from scalability issues. More recently, Backstrom and Leskovec [2011] presented a Supervised Random Walk (SRW) algorithm for link prediction that combines network structure and edge-attribute information, but this approach does not fully leverage node-attribute information, as it only incorporates node information for neighboring

¹<http://www.peakyou.com/> and <http://www.spokeo.com/>.

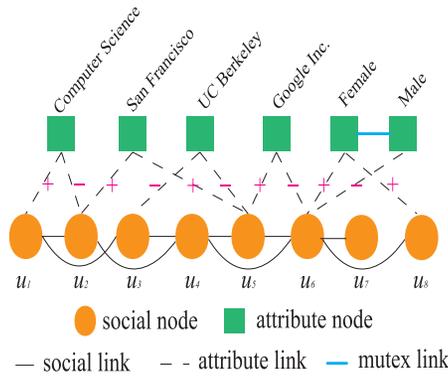


Fig. 1. Illustration of a social-attribute network (SAN). The link-prediction problem reduces to predicting social links, while the attribute-inference problem involves predicting attribute links.

nodes. For instance, SRW cannot take advantage of the common-node attribute San Francisco of u_2 and u_5 in Figure 1, since there is no edge between them.

Yin et al. [2010b, 2010a] proposed the use of a *social-attribute network* (SAN) to gracefully integrate network structure and node attributes in a scalable way. They focused on generalizing a Random Walk with Restart (RWwR) algorithm to the SAN model to predict links as well as to infer node attributes. In this article, we generalize several leading supervised and unsupervised link-prediction algorithms [Liben-Nowell and Kleinberg 2003; Hasan et al. 2006] to the SAN model in order to both predict links and infer missing attributes. We evaluate these algorithms on a novel, large-scale Google+ dataset and demonstrate performance improvement for each of them. Moreover, we make the novel observation that inferring attributes could help predict links, that is, link-prediction accuracy is further improved by first inferring missing node attributes.

2. PROBLEM DEFINITION

In our problem setting, we use an undirected² graph $G = (V, E)$ to represent a social network, where edges in E represent interactions between the $N = |V|$ nodes in V . In addition to network structure, we have categorical attributes for nodes. For instance, in the Google+ social network, nodes are users, edges represent friendship (or some other relationship) between users, and node attributes are derived from user profile information and include fields such as employer, school, and hometown. In this work, we restrict our focus to categorical variables, though in principle, other types of variables, for example, live chats, email messages, real-valued variables, etc., could be clustered into categorical variables via vector quantization, or directly discretized to categorical variables.

We use a binary representation for each categorical attribute. For example, various employers (e.g., Google, Intel, and Yahoo) and various schools (e.g., Berkeley, Stanford, and Yale) are each treated as separate binary attributes. Hence, for a specific social network, the number of distinct attributes M is finite (though M could be large). Attributes of a node u are then represented as an M -dimensional trinary column vector \vec{a}_u with the i th entry equal to 1 when u has the i th attribute (*positive attribute*), -1 when u does not have it (*negative attribute*), and 0 when it is unknown whether or not u has it (*missing attribute*). We denote by $A = [\vec{a}_1 \vec{a}_2 \cdots \vec{a}_N]$ the attribute matrix

²Our model and algorithms can also be generalized to directed graphs.

for all nodes. Note that certain attributes (e.g., female and male, age of 20 and 30) are mutually exclusive. Let L be the set of all pairs of mutually-exclusive attributes. This set constrains the attribute matrix A so that no column contains a 1 for two mutually-exclusive attributes.

We define the link-prediction problem as follows.

Definition 2.1 (Link-Prediction Problem). Let $T_i = (G_i, A_i, L_i)$ and $T_j = (G_j, A_j, L_j)$ be snapshots of a social network at times i and j . Then the link-prediction problem involves using T_i to predict the social network structure G_j . When $i < j$, new links are predicted. When $i > j$, missing links are predicted.

In this article, we work with three snapshots of the Google+ network crawled at three successive times, denoted $T_1 = (G_1, A_1, L_1)$, $T_2 = (G_2, A_2, L_2)$, and $T_3 = (G_3, A_3, L_3)$. To predict new links, we use various algorithms to solve the link-prediction problem with $i = 2$ and $j = 3$ and first learn any required hyperparameters by performing grid search on the link-prediction problem with $i = 1$ and $j = 2$. Similarly, to predict missing links, we solve the link-prediction problem with $i = 2$ and $j = 1$ and learn hyperparameters via grid search with $i = 3$ and $j = 2$.

For any given snapshot, several entries of A will be zero, corresponding to missing attributes. The attribute-inference problem, which involves only a single snapshot of the network, is defined as follows.

Definition 2.2 (Attribute-Inference Problem). Let $T = (G, A, L)$ be a snapshot of a social network. Then the attribute inference problem is to infer whether each zero entry of A corresponds to a positive or negative attribute, subject to the constraints listed in L .

Our goal is to design scalable algorithms leveraging both network structure and rich node attributes to address these problems for real-world, large-scale networks.

3. MODEL AND ALGORITHMS

3.1. Social-Attribute Network Model

The *social-attribute network* was first proposed by Yin et al. [2010a, 2010b]³ to predict links and infer attributes. However, their original model didn't consider negative and mutually-exclusive (mutex) attributes. In this section, we review this model and extend it to incorporate negative and mutex attributes.

Given a social network G with M distinct categorical attributes, an attribute matrix A and mutex attributes set L , we create an augmented network by adding M additional nodes to G , with each additional node corresponding to an attribute. For each node u in G with positive or negative attribute a , we create an undirected link between u and a in the augmented network. For each mutually exclusive attribute pair (a, b) , we create an undirected link between a and b . This augmented network is called the *social-attribute network* (SAN), since it includes the original social network interactions, relations between nodes and their attributes, and mutex links between attributes.

Nodes in the SAN model corresponding to nodes in G are called *social nodes*, and nodes representing attributes are called *attribute nodes*. Links between social nodes are called *social links*, and links between social nodes and attribute nodes are called *attribute links*. Attribute link (u, a) is a *positive attribute link* if a is a positive attribute of node u , and it is a *negative attribute link* otherwise. Links between mutually-exclusive attribute nodes are called *mutex links*. Intuitively, the SAN model explicitly describes

³Note that they name this model *Augmented Graph*. We call it *Social-Attribute Network* because it's more meaningful.

the sharing of attributes across social nodes as well as the mutual exclusion between attributes, as illustrated in the sample SAN model of Figure 1. Moreover, with the SAN model, the link-prediction problem reduces to predicting social links, and the attribute-inference problem involves predicting attribute links.

We also place weights on the various nodes and edges in the SAN model. These social-node weights could describe the individual tendencies (e.g., user activeness) of issuing social links, and the attribute-node weights could be used to balance the influence of social nodes versus attribute nodes. Furthermore, the edge weights of social links could describe their tie strengths, and the attribute-link weights could balance the influence of social links versus attribute links. We use $w(u)$ and $w(u, v)$ to denote the weight of node u and the weight of link (u, v) , respectively. Additionally, for a given social or attribute node u in the SAN model, we denote by $\Gamma_+(u)$ and $\Gamma_{s+}(u)$, respectively, the set of *all neighbors* and the set of *social neighbors* connected to u via social links or positive attribute links. We define $\Gamma_-(u)$ and $\Gamma_{s-}(u)$ in a similar fashion. This terminology will prove useful when we describe our generalization of leading link-prediction algorithms to the SAN model in the next section.

The fact that no social node can be linked to multiple mutex attributes is encoded in the *mutex property*, that is, there is no triangle consisting of a mutex link and two positive attribute links in any social-attribute network, which enforces a set of constraints for all attribute-inference algorithms.

In this work, we focus primarily on node attributes. However, we note that the SAN model can be naturally extended to incorporate *edge attributes*. Indeed, we can use a function (e.g., the logistic function) to map a given set of attributes for each edge (e.g., edge age) into the real-valued edge weights of the SAN model. The attributes-to-weight mapping function can be learned using an approach similar to the one proposed by Backstrom and Leskovec [2011].

3.2. Algorithms

Link-prediction algorithms typically compute a probabilistic score for each candidate link and subsequently rank these scores and choose the largest ones (up to some threshold) as putative new or missing links. In the following, we extend both unsupervised and supervised algorithms to the SAN model. Furthermore, we note that when predicting attribute links, the SAN model features a post-processing step whereby we change the lowest-ranked putative positive links violating the mutex property to negative links. Table I summarizes the various link-prediction and attribute-inference algorithms as well as their references.

3.2.1. Unsupervised Link and Attribute Inference. Liben-Nowell and Kleinberg [2003] provide a comprehensive survey of unsupervised link-prediction algorithms for social networks. These algorithms can be roughly divided into two categories: local-neighborhood-based algorithms and global-structure-based algorithms. In principle, all of the algorithms discussed in Liben-Nowell and Kleinberg [2003] can be generalized for the SAN model. In this work, we focus on representative algorithms from both categories, and we describe next how to generalize them to the SAN model to predict both social links and attribute links. We add the suffix ‘-SAN’ to each algorithm name to indicate its generalization to the SAN model. In our presentation of the unsupervised algorithms, we only consider positive attribute links, though many of these algorithms can be extended to signed networks [Symeonidis et al. 2010].

Common Neighbor (CN-SAN). is a local algorithm that computes a score for a candidate social or attribute link (u, v) as the sum of weights of u and v ’s common neighbors,

Table I. Summary of Various (a) Link-Prediction, (b) Attribute-Inference, and (c) Iterative-Link-and-Attribute-Inference Algorithms

| (a) Link Prediction | |
|--|--|
| Algorithm | References |
| Unsupervised | |
| Common Neighbor (CN) | [Liben-Nowell and Kleinberg 2003] |
| Common Neighbor (CN-SAN) | us |
| Adamic-Adar (AA) | [Adamic and Adar 2003], [Liben-Nowell and Kleinberg 2003] |
| Adamic-Adar (AA-SAN) | us |
| Low-rank Approximation (LRA) | [Liben-Nowell and Kleinberg 2003] |
| Low-rank Approximation (LRA-SAN) | us |
| CN + Low-rank Approximation (CN+LRA) | [Liben-Nowell and Kleinberg 2003] |
| CN + Low-rank Approximation (CN+LRA-SAN) | us |
| AA + Low-rank Approximation (AA+LRA) | us |
| AA + Low-rank Approximation (AA+LRA-SAN) | us |
| Random Walk with Restart (RWwR) | [Brin and Page 1998] [Pan et al. 2003] |
| Random Walk with Restart (RWwR-SAN) | [Yin et al. 2010b] |
| Supervised | |
| SLP-I | [Hasan et al. 2006], [Lichtenwalter et al. 2010] |
| SLP-II | [Hasan et al. 2006] |
| SLP-SAN-III/SLP-SAN-VI | us |

(b) Attribute Inference

| Algorithm | References |
|--|--------------------|
| Unsupervised | |
| Common Neighbor (CN-SAN) | us |
| Adamic-Adar (AA-SAN) | us |
| Low-rank Approximation (LRA-SAN) | us |
| CN + Low-rank Approximation (CN+LRA-SAN) | us |
| AA + Low-rank Approximation (AA+LRA-SAN) | us |
| Random Walk with Restart (RWwR-SAN) | [Yin et al. 2010b] |
| Supervised | |
| Supervised Attribute Inference (SAI-SAN) | us |

(c) Iterative Link and Attribute Inference

| Algorithm | References |
|--|------------|
| Iterative Link and Attribute Inference | us |

that is, $score(u, v) = \sum_{t \in \Gamma_+(u) \cap \Gamma_+(v)} w(t)$. Conventional CN only considers common social neighbors.

Adamic-Adar (AA-SAN) is also a local algorithm. For a candidate social link (u, v) , the AA-SAN score is

$$score(u, v) = \sum_{t \in \Gamma_+(u) \cap \Gamma_+(v)} \frac{w(t)}{\log |\Gamma_{s^+}(t)|}.$$

Conventional AA, initially proposed in Adamic and Adar [2003] to predict friendships on the Web and subsequently adapted by Liben-Nowell and Kleinberg [2003] to predict links in social networks, only considers common social neighbors. AA-SAN weights the importance of a common neighbor proportional to the inverse of the log of social

degree. Intuitively, we want to downweight the importance of neighbors that are either (i) social nodes that are social hubs or (ii) attribute nodes corresponding to attributes that are widespread across social nodes. Since in both cases, this weight depends on the social degree of a neighbor, the AA-SAN weight is derived based on social degree, rather than total degree.

In contrast, for a candidate attribute link (u, a) , the attribute degree of a common neighbor does influence the importance of the neighbor. For instance, consider two social nodes with the same social degree that are both common neighbors of nodes u and a . If the first of these social nodes has only two attribute neighbors while the second has 1,000 attribute neighbors, the importance of the former social node should be greater with respect to the candidate attribute link. Thus, AA-SAN computes the score for candidate attribute link (u, a) as

$$\text{score}(u, a) = \sum_{t \in \Gamma_{s^+}(u) \cap \Gamma_{s^+}(a)} \frac{w(t)}{\log |\Gamma_+(t)|}.$$

Low-Rank Approximation (LRA-SAN) takes advantage of global structure, in contrast to CN-SAN and AA-SAN. Denote X_S as the $N \times N$ weighted social-adjacency matrix, where the (u, v) th entry of X_S is $w(u, v)$ if (u, v) is a social link and zero otherwise. Similarly, let X_A be the $N \times M$ weighted attribute-adjacency matrix, where the (u, a) th entry of X_A is $w(u, a)$ if (u, a) is a positive attribute link and zero otherwise. We then obtain the weighted-adjacency matrix X for the SAN model by concatenating X_S and X_A , that is, $X = [X_S \ X_A]$. The LRA-SAN method assumes that a small number of latent factors (approximately) describes the social and attribute link strengths within X and attempts to extract these factors via low-rank approximation of X , denoted by \hat{X} . The LRA-SAN score for a candidate social or attribute link (u, t) is then simply \hat{X}_{ut} , or the (u, t) th entry of \hat{X} . LRA-SAN can be computed efficiently via truncated Singular Value Decomposition (SVD).

CN + Low-Rank Approximation (CN+LRA-SAN) is a mixture of local and global methods, as it first performs CN-SAN using a SAN model and then performs low-rank approximation on the resulting score matrix. After performing CN-SAN, let S_S be the resulting $N \times N$ score matrix for all social node pairs and S_A be the resulting $N \times M$ score matrix for all social-attribute node pairs. By virtue of the CN-SAN algorithm, note that S_S includes attribute information and S_A includes social interactions. CN+LRA-SAN then predicts social links by computing a low-rank approximation of S_S denoted \hat{S}_S , and each entry of \hat{S}_S is the predicted social-link score. Similarly, \hat{S}_A is a low-rank approximation of S_A , and each entry of \hat{S}_A is the predicted score for the corresponding attribute link.⁴

AA + Low-Rank Approximation (AA+LRA-SAN) is identical to CN+LRA-SAN but with the score matrices S_S and S_A generated via the AA-SAN algorithm.

Random Walk with Restart (RWwR-SAN) [Yin et al. 2010b] is a global algorithm. In the SAN model, a Random Walk with Restart [Brin and Page 1998; Pan et al. 2003] starting from u recursively walks to one of its neighbors t with probability proportional to the link weight $w(u, t)$ and returns to u with a fixed restart probability α . The probability $P_{u,v}$ is the stationary probability of node v in a random walk with restart initiated at u . In general, $P_{u,v} \neq P_{v,u}$. For a candidate social link (u, v) , we compute $P_{u,v}$

⁴An alternative method for combining CN-SAN and LRA-SAN under the SAN model that was not explored in this work involves defining $S = [S_S \ S_A]$, approximating S with \hat{S} , and using the (u, t) th entry of \hat{S} as a score for link (u, t) .

and $P_{v,u}$ and let $\text{score}(u, v) = (P_{u,v} + P_{v,u})/2$. Note that RWwR for link prediction in previous work [Liben-Nowell and Kleinberg 2003] computes these stationary probabilities based only on the social network. For a candidate attribute link (u, a) , RWwR-SAN only computes $P_{u,a}$, and $P_{u,a}$ is taken as the score of (u, a) .

We finally note that for predicting social links, if we set the weights of all attribute nodes and all attribute links to zero and we set the weights of all social nodes and social links to one, then all the algorithms just described reduce to their standard forms described in Liben-Nowell and Kleinberg [2003].⁵ In other words, we recover the link-prediction algorithms on pure social networks.

3.2.2. Supervised Link and Attribute Inference. Link prediction can be cast as a binary classification problem in which we first construct features for links and then use a classifier, such as SVMs or Logistic Regression. In contrast to unsupervised attribute inference, negative attribute links are needed in supervised attribute inference.

Supervised Link Prediction (SLP-SAN). For each link in our training set, we can extract a set of topological features F (e.g., CN, AA, etc.) computed from pure social networks and the similar features F_SAN computed from the corresponding social-attribute networks. We explored four feature combinations: (i) SLP-I uses only topological features F computed from social networks; (ii) SLP-II uses topological features F as well as an aggregate feature, that is, the number of common attributes of the two endpoints of a link; (iii) SLP-SAN-III uses topological features F_SAN ; and (iv) SLP-SAN-VI uses topological features F and F_SAN . SLP-SAN-III and SLP-SAN-VI contain the substring ‘SAN’ because they use features extracted from the SAN model. SLP-I and SLP-II are widely used in previous work [Hasan et al. 2006; Lichtenwalter et al. 2010].

Supervised Attribute Inference (SAI-SAN). Recall that attribute inference is transformed to attribute-link prediction with the SAN model. We can extract a set of topological features for each positive and negative attribute link. Moreover, the positive attribute links are taken as positive examples, while the negative attribute links are taken as negative examples. Hence, we can train a binary classifier for attribute links and then apply it to infer the missing attribute links.

3.2.3. Iterative Link and Attribute Inference. In many real-world networks, most node attributes are missing. Figure 2 shows the fraction of users as a function of the number of node attributes in the Google+ social network. From this figure, we see that roughly 70% of users have no observed node attributes. Hence, we will also investigate an iterative variant of the SAN model. We first infer the top attributes for users without any observed attributes. We then update the SAN model to include these predicted attributes and perform link prediction on the updated SAN model. This process can be performed for several iterations.

3.2.4. Scalability. The local unsupervised algorithms CN-SAN and AA-SAN are only concerned about the 2-hop neighborhoods, and thus are scalable. The low-rank-approximation-based unsupervised algorithms (i.e., LRA-SAN, CN+LRA-SAN, and AA+LRA-SAN) can be computed via truncated singular value decomposition (SVD), which was shown to be scalable [Talwalkar et al. 2008]. RWwR-SAN is based on a random walk with restart, which was also shown to be scalable [Tong et al. 2006]. Moreover, there are many papers/libraries making classifiers (e.g., SVM in our experiments.) scalable—for instance, Joachims [1999] and LIBLINEAR.⁶

⁵For LRA-SAN, this implies that X_A is an $N \times M$ matrix of zeros, so the truncated SVD of X is equivalent to that of X_S except for M zeros appended to the right singular vectors of X_S .

⁶<http://www.csie.ntu.edu.tw/~cjlin/liblinear/>.

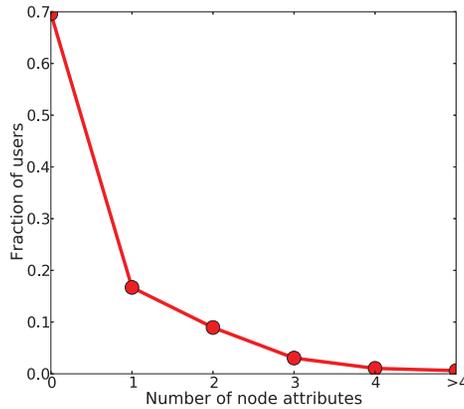


Fig. 2. The fraction of users as a function of the number of node attributes in the Google+ social network.

4. GOOGLE+ DATA

Google launched its new social network service named Google+ in early July 2011. We crawled three snapshots of the Google+ social network and its users' profiles on July 19, August 6, and September 19, 2011. They are denoted JUL, AUG, and SEP, respectively. We then pre-processed the data before conducting link-prediction and attribute-inference experiments.

Preprocessing Social Networks. In Google+, users divide their social connections into circles, such as a family circle and a friends circle. If user u is in v 's circle, then there is a directed edge (v, u) in the graph, and thus the Google+ dataset is a directed social graph. We converted this dataset into an undirected graph by only retaining edges (u, v) if both directed edges (u, v) and (v, u) exist in the original graph. We chose to adopt this filtering step for two reasons. (1) Bidirectional edges represent mutual friendships and hence represent a stronger type of relationship that is more likely to be useful when inferring users' attributes from their friends' attributes. (2) We reduce the influence of spammers who add people into their circles without those people adding them back. Spammers introduce fictitious directional edges into the social graph that adversely influence the performance of link-prediction algorithms.

Collecting Attribute Vocabulary. Google+ profiles include short entries about users, such as Occupation, Employment, Education, Places Lived, Gender, etc. We use Employment and Education to construct a vocabulary of attributes in this article, because previous work [Gong et al. 2012b] showed that these two attribute types are most powerful for forming links between people. We treat each distinct employer or school entity as a distinct attribute. Google+ has predefined employer and school entities, although users can still fill in their own defined entities. Due to users' changing privacy settings, some profiles in JUL are not found in AUG and SEP, so we use JUL to construct our attribute vocabulary. Specifically, from the profiles in JUL, we list all attributes and compute frequency of appearance for each attribute. Our attribute vocabulary is constructed by keeping attributes with frequency of at least 3.

Constructing Social-Attribute Networks. In order to demonstrate that the SAN model leverages node attributes well, we derived social-attribute networks in which each node has some positive attributes from the preceding Google+ social networks and attribute vocabulary. Specifically, for an attribute-frequency threshold k , we chose the largest connected social network from JUL such that each node has at least k distinct positive attributes. We also found the corresponding social networks consisting of these

Table II. Statistics of Social-Attribute Networks

| | #soci links | #all soci links | #soci nodes | #pos attri links | #attri nodes |
|------|-------------|-----------------|-------------|------------------|--------------|
| JUL4 | 7,062 | 7,062 | | | |
| AUG4 | 7,430 | 7,813 | 5,200 | 24,690 | 9,539 |
| SEP4 | 7,422 | 8,100 | | | |
| JUL2 | 287,906 | 287,906 | | | |
| AUG2 | 328,761 | 339,059 | 170,002 | 442,208 | 47,944 |
| SEP2 | 332,398 | 354,572 | | | |

nodes in snapshots AUG and SEP. Social-attribute networks were then constructed with the chosen social networks and the attributes of the nodes. Specifically, we chose $k = \{2, 4\}$ to construct six social-attribute networks whose statistics are shown in Table II. Each social-attribute network is named by concatenating the snapshot name and the attribute-frequency threshold. For example, ‘JUL4’ is the social-attribute network constructed using JUL and $k = 4$. These names are indicated in the first column of the table.

In the crawled raw networks, some social links in JUL_i are missing in AUG_i and SEP_i , where $i = 2, 4$. These links are missing due to one of two events occurring between the JUL and AUG or SEP snapshots: (1) users block other users, or (2) users set (part of) their circles to be publicly invisible, after which point they cannot be publicly crawled. These missed links provide ground truth labels for our experiments of predicting missing links. However, these missing links can alter estimates of network-level statistics and can have unexpected influences on link-prediction algorithms [Kossinets 2006]. Moreover, it is likely in practice that companies like Facebook and Google keep records of these missing links, and so it is reasonable to add these links back to AUG_i and SEP_i for our link-prediction experiments. The third column in Table II is the number of all social links after filling the missing links into AUG_i and SEP_i . The second column *#soci links* is used for experiments predicting missing links, and column *#all soci links* is used for the experiments predicting new links.

From these two columns, the number of new links or missing links can be easily computed. For example, if we use AUG2 as training data and SEP2 as testing data for link prediction, the number of new links is $354572 - 339059 = 15513$, which is computed with entries in column *#all soci links*. If we use AUG2 as training data and JUL2 as testing data in predicting missing links, the number of missing links is $339059 - 328761 = 10298$, which is computed with corresponding entries in column *#soci links* and *#all soci links*.

5. EXPERIMENTS

5.1. Experimental Setup

In our experiments, the main metric used is AUC, Area Under the Receiver Operating Characteristic (ROC) Curve, which is widely used in the machine learning and social network communities [Clauset et al. 2008; Backstrom and Leskovec 2011]. AUC is computed in the manner described in Hand and Till [2001], in which both positive and negative examples are required. In principle, we could use new links or missing links as positive examples and all non-existing links as negative examples. However, large-scale social networks tend to be very sparse, for example, the average degree is 4.17 in SEP2, and as a result, the number of non-existing links can be enormous, for example, SEP2 has around 2.9×10^{10} non-existing links. Hence, computing AUC using all non-existing links in large-scale networks is typically computationally infeasible. Moreover, the majority of new links in typical online social networks close triangles [Leskovec et al. 2008; Backstrom and Leskovec 2011], that is, are 2-hop links. For

instance, we find that 58% of the newly added links in Google+ are 2-hop links. We thus evaluate our large network experiments using 2-hop link data, as in Backstrom and Leskovec [2011], that is, new or missing 2-hop links are treated as positive examples, and non-existing 2-hop links are treated as negative examples.

In a social-attribute network, there are two categories of 2-hop links: (1) those with two endpoints sharing at least one common social node, and (2) those with two endpoints sharing only common attribute nodes. Local algorithms applied to the original social network are unable to predict 2-hop links in the second category. Thus, we evaluate only with respect to 2-hop links in the first category, so as not to give unfair advantage to algorithms running on the social-attribute network. To better understand whether the AUC performance computed on 2-hop links can be generalized to performance on any-hop links, we additionally compute AUC using any-hop links on the smaller Google+ networks.

In general, different nodes and links can have different weights in social-attribute networks, representing their relative importance in the network. In all of our experiments in this article, we set all weights to be one and leave it for future work to learn weights.

We compare our link-prediction algorithms with Supervised Random Walk (SRW) [Backstrom and Leskovec 2011], which leverages edge attributes, by transforming node attributes to edge attributes. Specifically, we compute the number of common attributes of the two endpoints of each existing link. As in Backstrom and Leskovec [2011], we also use the number of common neighbors as an edge attribute. We adopt the Wilcoxon-Mann-Whitney (WMW) loss function and logistic edge strength function in our implementations, as recommended in Backstrom and Leskovec [2011].

We compare our attribute-inference algorithms with two algorithms, BASELINE and LINK, introduced by Zheleva and Getoor [2009]. Using only node attributes, BASELINE first computes a marginal attribute distribution and then uses an attribute's probability as its score. LINK trains a classifier for each attribute by flattening nodes as the rows of the adjacency matrix of the social networks.⁷ Zheleva and Getoor [2009] found that LINK is the best algorithm when group memberships are not available.

We use SVM as our classifier in all supervised algorithms. For link prediction, we extract six topological features (CN-SAN, AA-SAN, LRA-SAN, CN+LRA-SAN, AA+LRA-SAN, and RWwR-SAN) from both pure social networks and social-attribute networks. Hence, SLP-I, SLP-II, SLP-SAN-III, and SLP-SAN-VI use 6, 7, 6, and 12 features, respectively. For attribute inference, we extract nine topological features for each attribute link. We adopt two ranks (detailed in Section 5.2.2) for each low-rank approximation-based algorithms, thus obtaining six features. The other three features are CN-SAN, AA-SAN, and RWwR-SAN. To account for the highly imbalanced class distribution of examples for supervised link prediction and attribute inference, we downsample negative examples so that we have equal numbers of positive and negative examples (techniques proposed in [Lichtenwalter et al. 2010; Doppa et al. 2010] could be used to further improve the performance).

We use the pattern *dataset1-dataset2* to denote a train-test or train-validation pair, with *dataset1* being a training dataset and *dataset2* being a testing or validation dataset. When conducting experiments to predict new links on the AUG_{*i*}-SEPi train-test pair, SRW, classifiers, and hyperparameters of global algorithms, that is, ranks in LRA-SAN, CN+LRA-SAN, and AA+LRA-SAN and the restart probability α in RWwR-SAN, are learned on the JUL_{*i*}-AUG_{*i*} train-validation pair. Similarly, when predicting

⁷The original LINK algorithm [Zheleva and Getoor 2009] trained a distinct classifier for each attribute type. In our setting, an attribute type (e.g., Education) can have multiple values, so we train a classifier for each binary attribute value.

Table III. Results for Predicting New Links

| (a) | | | (b) | | |
|------------|-----------|---------------|------------|-----------|---------------|
| Alg | w/o Attri | With Attri | Alg | w/o Attri | With Attri |
| Random | 0.5000 | 0.5000 | Random | 0.5000 | 0.5000 |
| CN-SAN | 0.6730 | 0.7315 | CN-SAN | 0.6936 | 0.7508 |
| AA-SAN | 0.7109 | 0.7476 | AA-SAN | 0.7638 | 0.7895 |
| LRA-SAN | 0.6003 | 0.6262 | LRA-SAN | 0.6410 | 0.6385 |
| CN+LRA-SAN | 0.6969 | 0.7671 | CN+LRA-SAN | 0.5642 | 0.6373 |
| AA+LRA-SAN | 0.7118 | 0.7471 | AA+LRA-SAN | 0.6032 | 0.6557 |
| RWwR-SAN | 0.6033 | 0.6143 | RWwR-SAN | 0.6788 | 0.6912 |

| (c) | | | (d) | |
|------------|-----------|---------------|-------------|-----------------------|
| Alg | w/o Attri | With Attri | Alg | AUC |
| Random | 0.5000 | 0.5000 | SLP-I | 0.9128(0.0140) |
| CN-SAN | 0.7482 | 0.8298 | SLP-II | 0.9580(0.0017) |
| AA-SAN | 0.7483 | 0.8324 | SLP-SAN-III | 0.9450(0.0007) |
| LRA-SAN | 0.8075 | 0.8237 | SLP-SAN-VI | 0.9706(0.0004) |
| CN+LRA-SAN | 0.7857 | 0.8651 | SRW | 0.9383 |
| AA+LRA-SAN | 0.8193 | 0.8552 | | |
| RWwR-SAN | 0.9363 | 0.9548 | | |

Note: (a) AUC of 2-hop new links on the train-test pair AUG4-SEP4. (b) AUC of 2-hop new links on the train-test pair AUG2-SEP2. (c) (d) AUC of any-hop new links on the train-test pair AUG4-SEP4. The numbers in parentheses are standard deviations.

missing links on train-test pair AUG_i-JUL_i , they are learned on train-validation pair SEP_i-AUG_i , where $i = 2, 4$.

The CN-SAN and AA-SAN algorithms are implemented in Python 2.7, while the RWwR-SAN algorithm and Supervised Random Walk (SRW) are implemented in Matlab, and all of them are run on a desktop with a 3.06GHz Intel Core i3 and 4GB of main memory. LRA-SAN, CN+LRA-SAN, and AA+LRA-SAN algorithms are implemented in Matlab and run on an x86-64 architecture using a single 2.60Ghz core and 30GB of main memory.

5.2. Experimental Results

In this section, we present evaluations of the algorithms on the Google+ dataset. We first show that incorporating attributes via the SAN model improves the performance of both unsupervised and supervised link-prediction algorithms. Then we demonstrate that inferring attributes via link-prediction algorithms within the SAN model achieves state-of-the-art performance. Finally, we show that by combining attribute inference and link prediction in an iterative fashion, we achieve even greater accuracy on the link-prediction task.

5.2.1. Link Prediction. To demonstrate the benefits of combining node attributes and network structure, we run the SAN-based link-prediction algorithms described in Section 3.2 both on the original social networks and on the corresponding social-attribute networks (recall that the SAN-based unsupervised algorithms reduce to standard unsupervised link-prediction algorithms when working solely with the original social networks).

Predicting New Links. Table III shows the AUC results of predicting new links for each of our datasets. We are able to draw a number of conclusions from these results. First, the SAN model improves every unsupervised learning algorithm on every dataset, save for LRA-SAN on AUG2-SEP2. The reason that LRA-SAN's performance

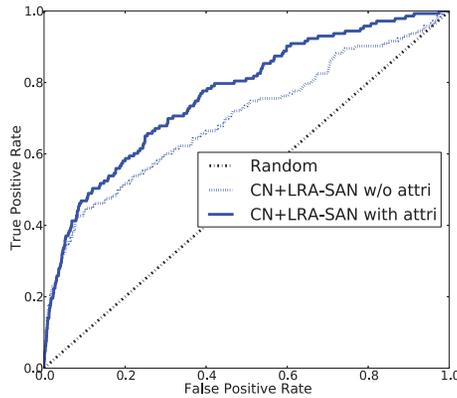


Fig. 3. ROC curves of the CN+LRA-SAN algorithm for predicting new links. AUG4-SEP4 is the train-test pair. JUL4-AUG4 is the train-validation pair.

decreases on AUG2-SEP2 could be that we searched the ranks in LRA-SAN up to 3,000 in the training and validation phase due to the limited computing resources available to us.⁸ Second, Table III(d) shows that attributes also improve supervised link-prediction performance, since SLP-SAN-VI, SLP-SAN-III, and SLP-II outperform SLP-I. Moreover, SLP-SAN-VI, which adopts features extracted from both social networks and social-attribute networks, achieves the best performance, thus demonstrating the power of the SAN model. Third, comparing RWwR-SAN in Table III(c) and SRW in Table III(d), we observe that the SAN model is better than SRW at leveraging node attributes, since RWwR-SAN with attributes outperforms SRW. This result is not surprising given that SRW is designed for edge attributes, and when transforming node attributes to edge attributes, we lose some information. For instance, as illustrated in Figure 1, nodes u_2 and u_5 share the attribute San Francisco. When transforming node attributes to edge attributes, this common attribute information is lost, since u_2 and u_5 are not linked.

Figure 3 shows the ROC curves of the CN+LRA-SAN algorithm. We see that the curve of CN+LRA-SAN with attributes dominates that of CN+LRA-SAN without attributes, demonstrating the power of the SAN model to effectively incorporate the additional predictive information of attributes. The results of other algorithms except LRA-SAN are similar and thus are not shown here.

Predicting Missing Links. Missing links can be divided into two categories: (1) links whose two endpoints have some social links in the training dataset, (2) links with at least one endpoint that has no social links in the training dataset. Category 1 corresponds to the scenarios where users block users or users set a part of their friend lists (e.g., family circles) to private. Category 2 corresponds to the scenario in which users hide their entire friend lists. Note that all 2-hop missing links belong to Category 1. In addition to performing experiments to show that the SAN model improves missing link prediction, we also perform experiments to explore which category of missing links is easier to predict. Table IV shows the results of predicting missing links on various datasets. As in the new-link prediction setting, the performance of every algorithm is improved by the SAN model, except for LRA-SAN on AUG4-JUL4 and RWwR-SAN on AUG4-JUL4 for 2-hop missing links.

⁸Note that LRA-SAN is scalable with more computing resources, which was shown by Talwalkar et al. [2008].

Table IV. Results for Predicting Missing Links

| (a) | | | (b) | | |
|------------|-----------|---------------|------------|-----------|---------------|
| Alg | w/o Attri | With Attri | Alg | w/o Attri | With Attri |
| Random | 0.5000 | 0.5000 | Random | 0.5000 | 0.5000 |
| CN-SAN | 0.7180 | 0.7925 | CN-SAN | 0.6938 | 0.7309 |
| AA-SAN | 0.7437 | 0.7697 | AA-SAN | 0.7633 | 0.7796 |
| LRA-SAN | 0.6569 | 0.6237 | LRA-SAN | 0.6044 | 0.6059 |
| CN+LRA-SAN | 0.7147 | 0.7986 | CN+LRA-SAN | 0.5816 | 0.6266 |
| AA+LRA-SAN | 0.7410 | 0.7668 | AA+LRA-SAN | 0.6212 | 0.6569 |
| RWwR-SAN | 0.5731 | 0.5676 | RWwR-SAN | 0.6595 | 0.6706 |

| (c) | | | (d) | | |
|------------|-----------|---------------|------------|-----------|---------------|
| Alg | w/o Attri | With Attri | Alg | w/o Attri | With Attri |
| Random | 0.5000 | 0.5000 | Random | 0.5000 | 0.5000 |
| CN-SAN | 0.5460 | 0.7012 | CN-SAN | 0.7329 | 0.7765 |
| AA-SAN | 0.5460 | 0.7033 | AA-SAN | 0.7330 | 0.7784 |
| LRA-SAN | 0.5495 | 0.6177 | LRA-SAN | 0.7316 | 0.7401 |
| CN+LRA-SAN | 0.5547 | 0.7048 | CN+LRA-SAN | 0.7515 | 0.7510 |
| AA+LRA-SAN | 0.5640 | 0.7325 | AA+LRA-SAN | 0.8104 | 0.8116 |
| RWwR-SAN | 0.2000 | 0.7619 | RWwR-SAN | 0.7797 | 0.8838 |

| (e) | | (f) | |
|-------------|-----------------------|-------------|-----------------------|
| Alg | AUC | Alg | AUC |
| SLP-I | 0.5453(0.0120) | SLP-I | 0.8023(0.0088) |
| SLP-II | 0.6991(0.0065) | SLP-II | 0.8403(0.0033) |
| SLP-SAN-III | 0.7161(0.0030) | SLP-SAN-III | 0.8620(0.0080) |
| SLP-SAN-VI | 0.8481(0.0022) | SLP-SAN-VI | 0.8854(0.0324) |

Note: (a) AUC of 2-hop missing links on the train-test pair AUG4-JUL4. (b) AUC of 2-hop missing links on the train-test pair AUG2-JUL2. (c)–(f) AUC of any-hop missing links on the train-test pair AUG4-JUL4. Missing links in both categories 1 and 2 are used in (c) and (e). Missing links in Category 1 are used in (d) and (f). The numbers in parentheses are standard deviations.

When comparing Tables IV(c) and IV(d) or Tables IV(e) and IV(f) we conclude that the missing links in Category 2 are harder to predict than those in Category 1. RWwR-SAN without attributes performs poorly when predicting any-hop missing links in both categories (as indicated by the entry with 0.2000 in Table IV(c)). This poor performance is due to the fact that RWwR-SAN without attributes assigns zero scores for all the missing links in Category 2 (positive examples) and positive scores for most non-existing links (negative examples), making many negative examples rank higher than positive examples and resulting in a very low AUC.

5.2.2. Attribute Inference. In this section, we focus on inferring attributes using the SAN model. In our next set of experiments in Section 5.2.3, we use the results of these attribute-inference algorithms to further improve link prediction, and the results of this iterative approach further validate the performance of the SAN model for attribute inference. Since the first step of iterative approach of Section 5.2.3 involves inferring the top attributes for each node, we employ an additional performance metric called Pre@ K in our attribute-inference experiments. Compared to AUC, Pre@ K better captures the quality of the top attribute predictions for each user. Specifically, for each sampled user, the top- K predicted attributes are selected, and (unnormalized) Pre@ K is then defined as the number of positive attributes selected divided by the number of sampled users. We address score ties in the manner described by McSherry and Najork [2008]. Since most Google+ users have a small number of attributes, we set $K = 2, 3, 4$ in our experiments.

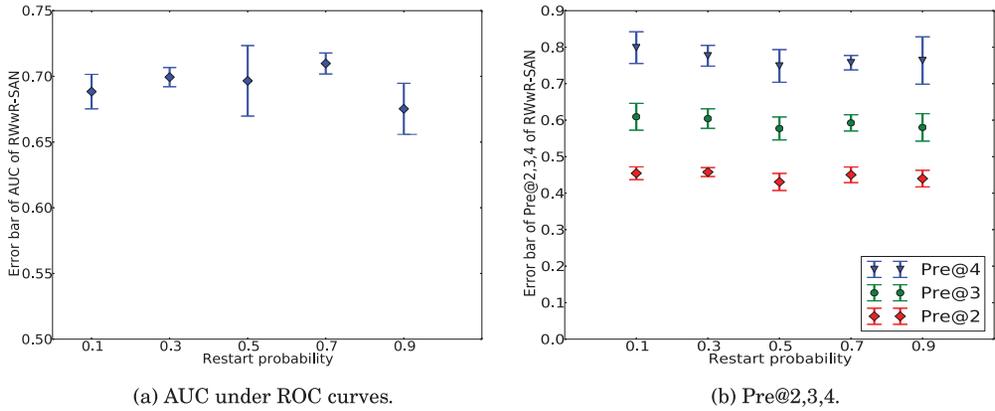


Fig. 4. Impact of the restart probability on the performance of RWwR-SAN for attribute inference on SEP4.

When evaluating algorithms for the inference of missing attributes, we require ground-truth data. In general, ground truth for node attributes is difficult to obtain, since it is often not possible to distinguish between negative and missing attributes. However, for most users, the number of attributes is quite small, and so we assume that users with many positive attributes have no missing attributes. Hence, we evaluate attribute inference on users that have at least four specified attributes, that is, we work with users in SEP4 and assume that each attribute link in SEP4 is either positive or negative.

In our experiment, we sample 10% of the users in SEP4 uniformly at random, remove their attribute links from SEP4, and evaluate the accuracy with which we can infer these users' attributes. All removed positive attribute links are viewed as positive examples, while all the negative attribute links of the sampled users are treated as negative examples. We run a variety of algorithms for attribute inference, and for each algorithm, we average the results over ten random trials. As previous noted, we evaluate the performance of attribute inference using both AUC and Pre@ K . Note that some attribute nodes in SEP4 only have one positive attribute link. As a result, if these positive attribute links are sampled as test data, it's hard to correctly infer them in the test phase, because there are no positive training examples for the corresponding attribute nodes. Therefore, we further remove the attribute nodes with only one positive attribute link and their corresponding attribute links from SEP4. Moreover, in the test phase, we only use the users among the sampled ones that have at least K attributes to compute the Pre@ K to avoid the influence of users with too few positive attributes.

For the low-rank approximation-based algorithms, that is, LRA-SAN, CN+LRA-SAN, and AA+LRA-SAN, we report results using two different ranks, 100 and 1,000, and indicate which was used by the number following the algorithm name in Figure 5. We choose these two small ranks for computational reasons and also based on the fact that low-rank approximation methods assume that a small number of latent factors (approximately) describes the social-attribute networks. Figure 4 shows the impact of the restart probability on the performance of RWwR-SAN. We find that $\alpha = 0.7$ achieves the best AUC and comparable Pre@ K . Thus we set the restart probability α to be 0.7.

Figure 5 shows the attribute-inference results for various algorithms. Several interesting observations can be made from this figure. First, under both metrics, all SAN-based algorithms perform better than BASELINE, save LRA100-SAN and LRA1000-SAN under Pre@2,3,4 metric, which indicates that the SAN model is good at leveraging

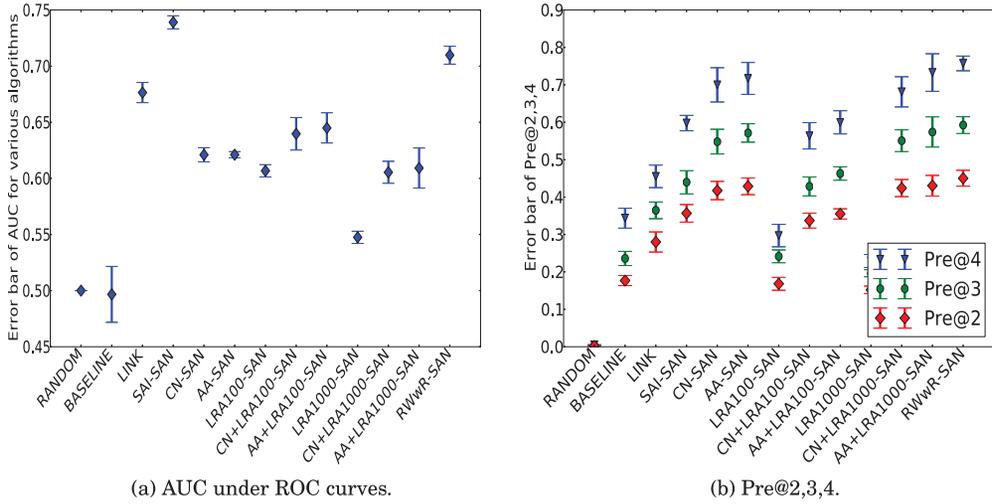


Fig. 5. Performance of various algorithms on attribute inference on SEP4.

network structure to infer missing attributes. Second, we find that AUC and $\text{Pre}@K$ provide inconsistent conclusions about relative algorithm performance. For instance, the mean AUC values suggest that SAI-SAN beats all other algorithms. However, several unsupervised algorithms outperform SAI-SAN with respect to $\text{Pre}@2,3,4$. The inconsistencies between the two metrics are expected, since AUC is a global measurement, while $\text{Pre}@K$ is a local one. Our SAI-SAN algorithm dominates LINK under both AUC and $\text{Pre}@2,3,4$ metrics, thus demonstrating the power of mapping attribute inference to link prediction with the SAN model.

5.2.3. Iterative Attribute and Link Inference. Section 5.2.1 demonstrated that knowledge of a user’s attributes can lead to significant improvements in link prediction. However, in real-world social networks like Google+, the vast majority of user attributes are missing (see Figure 2). To increase the realized benefits of social-attribute networks with few attributes, we first propose inferring missing attributes for each user whose attributes are missing and then performing link prediction on the inferred social-attribute networks. Recall that SAI-SAN achieves the best AUC, RWwR-SAN achieves the best $\text{Pre}@K$ in inferring attributes (see Figure 5), and AA-SAN achieves comparable $\text{Pre}@K$ results while being more scalable. Thus, in the following experiments, we use AA-SAN to first infer the top- K missing attributes for users, and subsequently perform link prediction using various methods.

In our experiments, when we are working on the pair *train-test*, we sample 10% of the users of *train* uniformly at random and remove their attributes. We then run three variants of link prediction algorithms: (i) without attributes, (ii) with only the remaining attributes, and (iii) with the remaining attributes along with the inferred attributes. The top-4 attributes are inferred for each sampled user by AA-SAN. We report the results averaged over ten trials. The hyperparameters of the global algorithms are the same as those in Section 5.2.1, which are learned from the corresponding train-validation pair.

Table V(a) shows the results of first inferring attributes and then predicting new links on the AUG4-SEP4 train-test pair. Table V(b) shows the results of first inferring attributes and then predicting missing links on the AUG4-JUL4 train-test pair. We see that the inferred attributes improve the performance of all algorithms except LRA-SAN on predicting missing links. Again, the bad performance of LRA-SAN could be

Table V. Results for Iteratively Inferring Attributes and Predicting Links

(a)

| Alg | w/o Attri | With Attri | With Inferred Attri |
|------------|-----------|----------------|------------------------|
| Random | 0.5000(0) | 0.5000(0) | 0.5000(0) |
| CN-SAN | 0.6730(0) | 0.7174(0.0077) | 0.7291(0.0063) |
| AA-SAN | 0.7109(0) | 0.7408(0.0063) | 0.7440(0.0026) |
| LRA-SAN | 0.6003(0) | 0.6274(0.0052) | 0.6320(0.0055) |
| CN+LRA-SAN | 0.6969(0) | 0.7497(0.0134) | 0.7534 (0.0084) |
| AA+LRA-SAN | 0.7111(0) | 0.7373(0.0050) | 0.7442(0.0032) |

(b)

| Alg | w/o Attri | With Attri | With Inferred Attri |
|------------|-----------|----------------|------------------------|
| Random | 0.5000(0) | 0.5000(0) | 0.5000(0) |
| CN-SAN | 0.7180(0) | 0.7780(0.0173) | 0.7856(0.0100) |
| AA-SAN | 0.7437(0) | 0.7626(0.0100) | 0.7661(0.0045) |
| LRA-SAN | 0.6569(0) | 0.6189(0.0105) | 0.6134(0.0157) |
| CN+LRA-SAN | 0.7147(0) | 0.7838(0.0256) | 0.7969 (0.0059) |
| AA+LRA-SAN | 0.7410(0) | 0.7591(0.0118) | 0.7673(0.0051) |

Note: (a) On the AUG4-SEP4 train-test pair. (b) On the AUG4-JUL4 train-test pair. Results are averaged over ten trials. The numbers in parentheses are standard deviations.

explained by the fact that we searched the ranks in LRA-SAN up to 3,000 in the training and validation phase due to the limited computing resources available to us. The AUCs obtained with inferred attributes for all other algorithms are very close to those obtained with all positive attributes, as shown in Table III(a), which further demonstrates that AA-SAN is an effective algorithm for attribute inference.

6. RELATED WORK

A wide range of link-prediction methods have been developed. For instance, models of complex networks, such as Preferential Attachment [Barabási and Albert 1999], SAN model [Gong et al. 2012b], and Hierarchical model [Clauset et al. 2008], can be viewed as models for predicting links. Liben-Nowell and Kleinberg [2003] surveyed a set of unsupervised link-prediction algorithms. Li et al. [2011] proposed Maximal Entropy Random Walk (MERW). Lichtenwalter et al. [2010] proposed the PropFlow algorithm which is similar to RWwR but more localized. However, none of these approaches leverage node-attribute information.

Link-prediction methods leveraging attribute information first appear in the relational learning community [Taskar et al. 2003; Miller et al. 2009; Bilgic et al. 2007; Yu et al. 2006]. However, these approaches suffer from scalability issues. For instance, the largest network tested in Taskar et al. [2003] has about $3K$ nodes. Recently, Backstrom and Leskovec [2011] proposed the Supervised Random Walk (SRW) algorithm to leverage edge attributes. However, SRW does not handle the scenario in which two nodes share common attributes (e.g., nodes u_2 and u_5 in Figure 1), but no edge already exists between them. Mapping link prediction to a classification problem [Hasan et al. 2006; Lichtenwalter et al. 2010; Doppa et al. 2010] is another way to incorporate attributes. We have shown that classifiers using features extracted from the SAN model perform very well. Yang et al. [2011] proposed jointly predicting links and propagating node interests (e.g., music interest). Their algorithm relies on the assumption that each node interest has a set of explicit attributes. As a result, their algorithm cannot be applied to our scenario in which it's hard (if possible) to extract explicit attributes for our node attributes.

Previous works [Rao et al. 2011, 2010] aim at inferring node attributes (e.g., ethnicity and political orientation) using supervised learning methods with features extracted from user names and user-generated texts. Zheleva and Getoor [2009] map attribute inference to a relational classification problem. They find that methods using group information achieve good results. These approaches are complementary to ours, since they use additional information apart from network structure and node attributes. In this article, we transform the attribute-inference problem into a link-prediction problem with the SAN model. Therefore, any link-prediction algorithm can be used to infer missing attributes. More importantly, we demonstrate that attribute inference can in turn help link prediction with the SAN model.

7. CONCLUSION AND FUTURE WORK

We comprehensively evaluate the social-attribute network (SAN) model in terms of link prediction and attribute inference. More specifically, we adapt several representative unsupervised and supervised link-prediction algorithms to the SAN model to both predict links and infer attributes. Our evaluation with a large-scale novel Google+ network dataset demonstrates performance improvement for each of these generalized algorithms on both link prediction and attribute inference. Moreover, we demonstrate further improvement of link-prediction accuracy by using the SAN model in an iterative fashion, first to infer missing attributes and subsequently to predict links. Interesting avenues for future research include devising an iterative algorithm that alternates between attribute and link prediction, learning node and edge weights in the SAN model, and incorporating edge attributes, negative node attributes, and mutex edges into large-scale experiments.

ACKNOWLEDGMENTS

We would like to thank Di Wang, Satish Rao, Mario Frank, Kurt Thomas, and Shobha Venkataraman for insightful feedback.

REFERENCES

- L. A. Adamic and E. Adar. 2003. Friends and neighbors on the web. *Social Netw.* 25, 3, 211–230.
- L. Backstrom and J. Leskovec. 2011. Supervised random walks: Predicting and recommending links in social networks. In *Proceedings of the WSDM Conference*.
- A.-L. Barabási and R. Albert. 1999. Emergence of scaling in random networks. *Science* 286, 509–512.
- S. Bartunov, A. Korshunov, S.-T. Park, W. Ryu, and H. Lee. 2012. Joint link-attribute user identity resolution in online social networks. In *Proceedings of the Workshop on Social Network Mining and Analysis (SNA-KDD)*.
- M. Bilgic, G. Namata, and L. Getoor. 2007. Combining collective classification and link prediction. In *Proceedings of the ICDM Workshops*. IEEE Computer Society, 381–386.
- S. Brin and L. Page. 1998. The anatomy of a large-scale hypertextual Web search engine. *Comput. Netw. ISDN Syst.* 30, 1–7, 107–117.
- A. Clauset, C. Moore, and M. E. J. Newman. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 453, 7191, 98–101.
- J. R. Doppa, J. Yu, P. Tadepalli, and L. Getoor. 2010. Learning algorithms for link prediction based on chance constraints. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*. 344–360.
- T. L. Fond and J. Neville. 2011. Randomization tests for distinguishing social influence and homophily effects. In *Proceedings of the World Wide Web Conference (WWW)*. ACM, New York, NY, 601–610.
- N. Z. Gong, A. Talwalkar, L. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. Shi, and D. Song. 2012a. Jointly predicting links and inferring attributes using a social-attribute network (san). In *Proceedings of the Workshop on Social Network Mining and Analysis (SNA-KDD)*.
- N. Z. Gong, W. Xu, L. Huang, P. Mittal, E. Stefanov, V. Sekar, and D. Song. 2012b. Evolution of social-attribute networks: Measurements, modeling, and implications using google+. In *Proceedings of the Internet Measurement Conference (IMC)*.

- D. J. Hand and R. J. Till. 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. *Machine Learn.* 45, 171–186.
- M. A. Hasan, V. Chaoji, S. Salem, and M. Zaki. 2006. Link prediction using supervised learning. In *Proceedings of the SIAM Workshop on Link Analysis, Counterterrorism and Security*.
- T. Joachims. 1999. Making large-scale SVM learning practical. In *Advances in Kernel Methods - Support Vector Learning*, MIT Press, 169–184.
- M. Kim and J. Leskovec. 2011. Modeling social networks with node attributes using the multiplicative attribute graph model. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI)*.
- G. Kossinets. 2006. Effects of missing data in social networks. *Social Netw.* 28, 247–268.
- G. Kossinets and D. Watts. 2006. Empirical analysis of an evolving social network. *Science* 311, 5757, 88–90.
- R. Kumar, J. Novak, P. Raghavan, and A. Tomkin. 2004. Structure and evolution of blogspace. *Communi. ACM* 47, 12, 35–39.
- J. Leskovec, L. Backstrom, R. Kumar, and A. Tomkins. 2008. Microscopic evolution of social networks. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 462–470.
- R.-H. Li, J. X. Yu, and J. Liu. 2011. Link prediction: The Power of maximal entropy random walk. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*.
- D. Liben-Nowell and J. Kleinberg. 2003. The link prediction problem for social networks. In *Proceedings of the ACM Conference on Information and Knowledge Management (CIKM)*. 556–559.
- R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. 2010. New perspectives and methods in link prediction. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- F. McSherry and M. Najork. 2008. Computing information retrieval performance measures efficiently in the presence of tied scores. In *Proceedings of the European Conference on Information Retrieval (ECIR)*.
- P. Melville and V. Sindhvani. 2010. Recommender systems. In *Encyclopedia of Machine Learning*. Springer.
- A. K. Menon and C. Elkan. 2011. Link prediction via matrix factorization. In *Proceedings of the European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD)*.
- K. T. Miller, T. L. Griffiths, and M. I. Jordan. 2009. Nonparametric latent feature models for link prediction. In *Proceedings of the Neural Information Processing Systems Conference (NIPS)*.
- J.-Y. Pan, H.-J. Yang, C. Faloutsos, and P. Duygulu. 2003. Automatic multimedia cross-modal correlation discovery. In *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*.
- D. Rao, M. Paul, C. Fink, D. Yarowsky, T. Oates, and G. Coppersmith. 2011. Hierarchical bayesian models for latent attribute detection in social networks. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.
- D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta. 2010. Classifying latent user attributes in twitter. In *Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents (SMUC)*. ACM, New York, NY, 37–44.
- J. Scripps, P.-N. Tan, F. Chen, and A.-H. Esfahanian. 2009. A matrix alignment approach for collective classification. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.
- A. P. Singh and G. J. Gordon. 2008. Relational learning via collective matrix factorization. In *Proceedings of the KDD*.
- P. Symeonidis, E. Tiakas, and Y. Manolopoulos. 2010. Transitive node similarity for link prediction in social networks with positive and negative links. In *Proceedings of the ACM Recommender System Conference (RecSys)*.
- A. Talwalkar, S. Kumar, and H. Rowley. 2008. Large-scale manifold learning. In *Proceedings of the CVPR*. 273–297.
- B. Taskar, M.-F. Wong, P. Abbeel, and D. Koller. 2003. Link prediction in relational data. In *Proceedings of the NIPS*.
- H. Tong, C. Faloutsos, and J.-Y. Pan. 2006. Fast random walk with restart and its applications. In *Proceedings of the ICDM*.
- S. H. Yang, B. Long, A. Smola, N. Sadagopan, Z. Zheng, and H. Zha. 2011. Like like alike—joint friendship and interest propagation in social networks. In *Proceedings of the WWW*. 537–546.
- Z. Yin, M. Gupta, T. Weninger, and J. Han. 2010a. LINKREC: A unified framework for link recommendation with user attributes and graph structure. In *Proceedings of the WWW*. 1211–1212.

- Z. Yin, M. Gupta, T. Wenginger, and J. Han. 2010b. A unified framework for link recommendation using random walks. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*.
- K. Yu, W. Chu, S. Yu, V. Tresp, and Z. Xu. 2006. Stochastic relational models for discriminative link prediction. In *Proceedings of the NIPS*.
- E. Zheleva and L. Getoor. 2009. To join or not to join: The illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the WWW*.

Received October 2012; revised February 2013; accepted March 2013