Network Flows and the Link Prediction Problem

Kanika Narang
Indraprastha Institute of Information Technology Delhi,India
kanika08029@iiitd.ac.in

Kristina Lerman
USC Information Sciences Institute, CA, USA
lerman@isi.edu

Ponnurangam Kumaraguru
Indraprastha Institute of Information Technology Delhi, India
pk@iiitd.ac.in

ABSTRACT

Link prediction is used by many applications to recommend new products or social connections to people. Link prediction leverages information in network structure to identify missing links or predict which new one will form in the future. Recent research has provided a theoretical justification for the success of some popular link prediction heuristics, such as the number of common neighbors and the Adamic-Adar score, by showing that they estimate the distance between nodes in some latent feature space. In this paper we examine the link prediction task from the novel perspective of network flows. We show that how easily two nodes can interact with or influence each other depends not only on their position in the network, but also on the nature of the flow that mediates interactions between them. We show that different types of flows lead to different notions of network proximity, some of which are mathematically equivalent to existing link prediction heuristics. We measure the performance of different heuristics on the missing link prediction task in a variety of real-world social, technological and biological networks. We show that heuristics based on a random walk-type processes outperform the popular Adamic-Adar and the number of common neighbors heuristics in many networks.

1. INTRODUCTION

Link prediction is a the heart of many graph mining and network analysis applications, including recommendations for new products or social connections. The link prediction task can be stated as follows: given a network, or a graph, predict what edges will form between nodes in the future. Alternatively, in domains where data collection is costly and the resulting graphs are noisy and incomplete, link prediction can be used to identify unobserved edges. A variety of heuristics have been proposed for the link prediction task, including those based on various notions of network proximity, such as neighborhood overlap [8], the Adamic-Adar score [1], which weighs the contribution of each common neighbor by the inverse of the logarithm of its degree, and number of paths connecting the two nodes. Several studies have compared the effectiveness of these heuristics on link prediction task in real-world networks [17, 20, 14, 27, 31, 11, 3, 19]. Generally, simple local measures, such as the number of common neighbors and the Adamic-Adar score, work best in predicting new or missing links [17, 20]. Recently, Sarkar et al. [25] explained why these heuristics work so well. They envision nodes residing in some latent space, where more similar nodes are closer to one other than dissimilar nodes, and prove that link prediction heuristics, such as the number of common neighbors, estimate distance between nodes in the latent space. In this view, specific network topology is a realization of the positions of the nodes in the latent space; therefore, network proximity estimates the true distance between nodes.

We argue that link prediction heuristic should take into account not only how close two nodes are in a network, but also their ability to exchange information or to influence each other. This is determined by the nature of the *flow* taking place on the network, i.e., the process by which information or influence is transmitted from one node to another. The flow shapes our view of who the important nodes are [4, 15], as well as the community structure of the network [5, 16]. In this paper we show that network flows also affect the notion of network proximity. Consider a social network in which people pass around a book. Each person chooses one friend and sends her the book. After finishing the book, the friend will mail the book to one of her friends. Used-good circulation, money exchange, and Web surfing are all examples of such one-to-one interactions, which can be modeled as a random walk. In the book exchange network, two individuals can be considered to be socially close even if they don't know each other, if a book mailed by one is often received by the other, and vice versa. In other words, the probability that a random walk originating at one node reaches the other measures the proximity of two nodes in the network. Heuristics based on the random walk-based proximity were used by [14, 3] in the link prediction task.

The spread of a virus through a population, on the other hand, cannot be modeled as a random walk. In an epidemic, rather than picking one neighbor, a node attempts to infect all neighbors. As we show in this paper, network proximity in this case can be measured by the number of common neighbors, one of the most common link prediction heuristics. Still other network flows are possible. Consider a flow in which a node's capacity to receive incoming messages is limited by its bandwidth. As a consequence, the more incoming connections (in-links) a node has, the less likely it is to receive a message from an arbitrary connection, e.g., because it has already reached the limit of its capacity by processing other incoming messages. This alters the char-

The 7th SNA-KDD Workshop '13 (SNA-KDD'13), August 11, 2013, Chicago, United States.

acter of the flow and leads to novel measures of network proximity.

In this paper, we define a flow-based network proximity and introduce several novel proximity measures based on different types of flow (Section 3). We relate well known link prediction heuristics to these proximity measures and introduce new ones that have not previously been considered in literature. Our approach unifies link prediction heuristics by viewing them as instances of network proximity under different network flows. In Section 4 we compare the performance of the new and existing heuristics on the missing link prediction task in a variety of real-world social, technological and biological networks. Our work adds a new dimension to this popular problem by connecting different link prediction heuristics to dynamic flows taking place on the network.

2. BACKGROUND

We represent a network as a directed, unweighed graph $G = (V, E)$ with V nodes and E edges. An example of a directed graph is shown in Fig. 1, where an edge direction indicates the direction of the flow of information or influence. Nodes receive messages from their in-neighbors and send messages to their out-neighbors. Such a graph could represent the Web, with nodes as Web pages and edges as hyperlinks between pages, or airports linked by direct routes, etc.

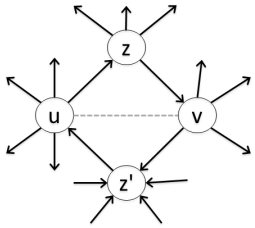

Figure 1: Example of a directed graph.

The adjacency matrix of the graph is defined as: $A(u, v) = 1$ if $(u, v) \in E$; otherwise, $A(u, v) = 0$. The set of out-neighbors of u is $\Gamma_{\text{out}}(u) = \{v \in V | (u, v) \in E\}$, and the out-degree of u is

$$d_{\text{out}}(u) = \sum_{v \in V} A(u, v) = |\Gamma_{\text{out}}(u)|.$$

Similarly, $\Gamma_{\text{in}}(u)$ represents the set of in-neighbors of u, and $d_{\text{in}}(u)$ is the in-degree of u. The total degree of the node is $d(u) = d_{\text{out}}(u) + d_{\text{in}}(u)$. In undirected graph, $A(u, v) = A(v, u)$, and the neighborhood of u, $\Gamma(u)$, consists of nodes that are connected to u: $\Gamma(u) = \{v \in V | (u, v) \in E\}$.

Intuitively, the closer the nodes are in a network, the easier it is for them to exchange information or to influence one another. Link prediction studies leveraged this intuition by using measures of network proximity to predict new or missing links. Some of the measures used in previous studies [17, 20, 31] are listed in Table 1 and include the number of common neighbors (CN), fraction of common neighbors, or Jaccard (JC) coefficient, Adamic-Adar (AA) score [1], which weighs each common neighbor by the inverse of the logarithm of its degree, and resource allocation (RA) measure [31], which weighs each common neighbor by the inverse of its degree.

Table 1: Heuristics used in link prediction applications. Popular existing link prediction heuristics appear above the double line: number of common neighbors, Jaccard and Adamic-Adar score, and resource allocation. Below the double line are link prediction heuristics introduced in this paper.

name	symbol
common neighbors	CN
Jaccard score	JC
Adamic-Adar	AA
resource allocation	RA
conservative (random walk)	CS
limited-bandwidth conservative	lCS
non-conservative (epidemic)	NC
limited-bandwidth non-conservative	lNC
hybrid conservative	hCS
hybrid limited-bandwidth conservative	$hlCS$
hybrid non-conservative	lNC
hybrid limited-bandwidth non-conservative	$hlNC$

Sarkar et al. [25] explained why these simple heuristics work well in the link prediction task. They imagine nodes embedded in a metric latent space, each dimension of which corresponds to some feature that nodes share. Nodes that are close to each other in the latent space are more similar than more distant nodes. They showed that some of the popular link prediction heuristics, such as the number of common neighbors, provide a good bound for the distance between nodes. However, their work did not account for the nature of the flow between nodes and its effect on how readily they can exchange information or influence each other.

3. NETWORKS FLOWS AND PROXIMITY

The likelihood that information or influence will reach the target node from the source node depends not only on network topology, but also on *how information flows on the network*. The flow is a stochastic dynamic process whose transitions are mediated by the interactions between nodes. Consider a graph of hyperlinked Web pages, for example. The process of browsing this graph is best described as a random walk. At each page, a Web surfer picks one of the hyperlinks to a neighbor of that page in the Web graph and navigates to it. The flows in the air transportation network and one-to-one social interactions, such as book exchange, can also be modeled as a random walk. Not all flows, however, can be modeled by a random walk, and different types of interaction lead to different notions of proximity even in the same network. Below we discuss several different types of network flows and define proximity measures for each type of flow.

3.1 Random walks

A random walk is a stochastic process that starts at some node and transitions to a randomly chosen out-neighbor of the node. Random walks are used to model a variety of physical processes based on diffusion, but also social processes, such as Web surfing, money and used goods exchange. These can be modeled as one-to-one interactions, since each person must first choose one out-neighbor to interact with. Because random walks conserve the probability distribution of the stochastic process, we call network flows based on random walks *conservative*.

Random walk-based proximity measures compute the probability a random walk starting at source node u will reach the target node v through any path in the graph [14]. A walker starting at node u can reach v only through a common neighbor z. While there could be longer paths that connect u to v, a local measure considers only paths of length two that go through intermediate nodes such as z or z' in Fig. 1. A random walker moving from u to v first needs to pick an edge that will take it from u to z, which it will do with probability $1/d_{\text{out}}(u)$. Then it has to pick an edge that will take it from z to v, which it will do with probability $1/d_{\text{out}}(z)$. Since we are interested in an edge-based measure, either u or v could be the source of the message. Therefore, we have to make the measure symmetric by considering flows from either direction. Symmetrizing, we obtain a random walk-based proximity measure, that we refer to as *conservative* measure, which gives the probability a random walk will reach u from v or vice versa through paths of length two:

$$CS = \frac{1}{2}\Big[\sum_{z\in\Delta}\frac{1}{d_{\text{out}}(u)d_{\text{out}}(z)} + \sum_{z\in\Delta'}\frac{1}{d_{\text{out}}(v)d_{\text{out}}(z)}\Big]. \quad (1)$$

This measure is defined in terms of the directed neighborhoods of the source u and target v nodes:

$$\begin{aligned}\Delta &= \Gamma_{\text{out}}(u)\cap\Gamma_{\text{in}}(v)\\ \Delta' &= \Gamma_{\text{in}}(u)\cap\Gamma_{\text{out}}(v).\end{aligned}$$

In an undirected graph, neighborhood overlap is defined as $\bar{\Delta} = \Gamma(u)\cap\Gamma(v)$, and conservative proximity measure reduces to

$$\overline{CS} = \frac{1}{2}\Big[\frac{1}{d(u)} + \frac{1}{d(v)}\Big]\sum_{z\in\bar{\Delta}}\frac{1}{d(z)}. \quad (2)$$

Like the Adamic-Adar score [1] and resource allocation measure [31], conservative proximity downweighs the neighbors' contributions to proximity by their degree.

3.2 Bandwidth-limited random walks

In the discussion above, we assumed that the target node has unlimited capacity to receive incoming messages. This may not always be true. Suppose a Web server can process a limited number of http requests, in an extreme case only one. Then the probability that a Web surfer starting at page u will reach z depends on whether the Web server in charge of page z is available to process the incoming page request. If it is already processing another request received through an incident hyperlink, it will reject the current request. If walkers arrive at random, the probability a walker will successfully transition to the target node is inversely proportional to the number of target node's incoming edges.

In Fig. 1, node z has one incoming edge. A walker starting at u will successfully transition to z. In contrast, z' has five incoming edges; therefore, a walker attempting a transition to z' will be successful one fifth of the time.

Nodes' limited bandwidth alters the flow of messages on the network. Now, in order for a message to get from u to z, not only must u pick an edge that will get the message to z, but z must also have the bandwidth to receive that message. We model the effect of limited bandwidth by weighing the probability the node will receive a message by the inverse of its in-degree $1/d_{in}(z)$. The proximity measure for nodes interacting via limited-bandwidth random walks is:

$$\begin{aligned}lCS &= \frac{1}{2}\Big[\sum_{z\in\Delta}\frac{1}{d_{\text{out}}(u)d_{\text{in}}(z)d_{\text{out}}(z)d_{\text{in}}(v)}\\ &\quad + \sum_{z\in\Delta'}\frac{1}{d_{\text{out}}(v)d_{\text{in}}(z)d_{\text{out}}(z)d_{\text{in}}(u)}\Big].\end{aligned}$$

3.3 Epidemics

Now imagine that information flows on a network not via one-to-one interactions, which can be modeled as a random walk, but via one-to-many broadcasts, which can be modeled as an epidemic process. Epidemics are used to describe many types of social processes, including the spread of a disease or innovation in a social network. In contrast to a random walk, in an epidemic, a node attempts to infect, or broadcast a message to, all of its out-neighbors. Because epidemics do not have a conserved quantity associated with it, we call such flows *non-conservative*.

In Fig. 1, for a message to get from node u to v via epidemic interactions, first u broadcasts it to its out-neighbors, including z, and then z broadcasts it to its own out-neighbors. Probability of the message being transmitted from one node to another is one. Therefore, symmetrized epidemic-based proximity measure is:

$$NC = \frac{1}{2}\Big[\sum_{z\in\Delta}1 + \sum_{z\in\Delta'}1\Big] = \frac{1}{2}\big[|\Delta| + |\Delta'|\big]. \quad (3)$$

This measure counts the expected number of times a message is received and is identical to the neighborhood overlap measure CN. While this measure was originally motivated by the intuition that when people have many friends in common, they are more likely to attend the same events and be in the same community, therefore, considered "close" [8], our work shows that it also can be derived from the principles of one-to-many interactions, of which social interactions are a prime example.

3.4 Bandwidth-limited epidemics

A node's limited capacity to receive incoming messages can also affect dynamics of an epidemic. Consider, as an example, the spread of information in a social network, which is often modeled as an epidemic process. However, people have finite attention [13], or effort they are willing to spend on some task, including responding to messages they receive from their friends. Not only is the attention limited, but people must also divide it among all friends [10]. Hence, a person's probability to respond to a message from a friend decreases with the number of friends she follows [10].

For simplicity, we assume that a node's attention, or bandwidth, is distributed uniformly over all in-neighbors. We model the effect of limited bandwidth by weighing the probability the node will receive a message by the inverse of its in-degree $1/d_{in}(z)$. In an example in Fig. 1, node z has one in-neighbor, and is able to receive all messages that it sends. In contrast, node z' has five in-neighbors, and will receive a message that an in-neighbor sends only 20% of the time.

Nodes' limited bandwidth alters the nature of epidemic flow on the network, and how easily nodes can communicate or influence each other. The resulting proximity between nodes can be written as:

$$lNC = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{d_{\text{in}}(z)d_{\text{in}}(v)} + \sum_{z\in\Delta'}\frac{1}{d_{\text{in}}(z)d_{\text{in}}(u)}\right]. \quad (4)$$

In undirected graphs, this reduces to:

$$\overline{lNC} = \frac{1}{2}\left[\frac{1}{d(u)} + \frac{1}{d(v)}\right]\sum_{z\in\Gamma(u)\cap\Gamma(v)}\frac{1}{d(z)},$$

which is identical to conservative proximity (CS) in undirected networks (Eq. 2).

3.5 Hybrid flows

In the examples above we considered pure flows, in which all nodes interacted using the same rules. This need not always be the case: in some cases a flow could be hybrid, composed of flows of different types. While a great variety of hybrid flows are possible, the most useful ones to consider are those in which the source and target nodes broadcast message via epidemic diffusion and do not have any bandwidth limitations, while their neighbors continue to be bound by the rules of the respective flow. One useful feature of this type of hybrid flow is that the resulting proximity measure does not contain the properties of the source and target nodes, e.g., their in-degree, and is, therefore, symmetric. Some of the more popular link prediction heuristics, such as Adamic-Adar score and RA, are independent of the properties of the end point nodes and only consider the properties of the neighboring nodes.

Hybrid conservative proximity measures the likelihood a message sent by the source node will be received by the target node (and vice versa) when the source node broadcasts the message, which is relayed to the target node by the common neighbors via a random walk. The probability that a common neighbor z forwards the message to the target node is proportional to inverse of z's out-degree. Hence, hybrid conservative proximity is:

$$hCS = \frac{1}{2}\Big[\sum_{z\in\Delta}\frac{1}{d_{\text{out}}(z)} + \sum_{z\in\Delta'}\frac{1}{d_{\text{out}}(z)}\Big]. \quad (5)$$

In an undirected network, this becomes

$$\overline{hCS} = \sum_{z\in\bar{\Delta}}\frac{1}{d(z)}.$$

This measure is identical to the resource allocation measure RA shown by Zhou et al. [31] to give good performance on the task of predicting missing links in several real-world networks, including the electric power grid, router-level Internet graph, and US air transportation network.

Similarly, proximity measures for other types of hybrid flows that do not include the in- and out-degree of the source and target nodes can be written as:

$$hlCS = \frac{1}{2}\Big[\sum_{z\in\Delta}\frac{1}{d_{\text{in}}(z)d_{\text{out}}(z)} + \sum_{z\in\Delta'}\frac{1}{d_{\text{in}}(z)d_{\text{out}}(z)}\Big] \quad (6)$$

$$hNC = \frac{1}{2}\left[|\Delta| + |\Delta'|\right] \quad (7)$$

$$hlNC = \frac{1}{2}\left[\sum_{z\in\Delta}\frac{1}{d_{\text{in}}(z)} + \sum_{z\in\Delta'}\frac{1}{d_{\text{in}}(z)}\right]. \quad (8)$$

Note that non-conservative proximity is the same as the hybrid non-conservative proximity. Also, in undirected networks, there is no distinction between hybrid limited-bandwidth non-conservative proximity and hybrid conservative proximity:

$$\overline{hlNC} = \sum_{z\in\bar{\Delta}}\frac{1}{d(z)} = \overline{hCS} = RA.$$

4. EMPIRICAL STUDY

Missing link prediction task has been used to evaluate link prediction heuristics on networks for which temporal information about creation of edges is not available [31, 20]. This task has other applications in network analysis, for example, predicting edges in partially observed networks for which obtaining complete data may be infeasible or prohibitively expensive, for example, criminal networks.

4.1 Missing Link Prediction Task

Each trial of the missing link prediction task is composed of the following steps. First, we randomly remove 10% of all edges and assign them to the test set E_{test}. The remaining 90% of links comprise the training set, the graph $G_{train} = (V, E_{train})$. While removing the edges, care is taken to ensure that the graph remains connected, i.e., deleting an edge will not cause the node to become disconnected from the rest of the graph and node should have a minimum degree of 2 after removal of an edge. We then compute network proximity using a given link prediction heuristic for all pairs of nodes $|V \times V - E_{train}|$ and rank them in decreasing order. We take the top-M predicted edges $E_{predicted}(M)$ and score the prediction based on how many of them are correct: $correct(M)=|E_{predicted}(M) \bigcap E_{test}|$.

We evaluated the performance of metrics in the link prediction task using the Receiver Operating Characteristic (ROC) Curve, which has been extensively used in data mining and machine learning research.The ROC curve is plotted using the fraction of true positives (TP Rate) vs. fraction of false positives (FP Rate) at different values of experimental parameter. For our experiment, we computed the two ratios at varying values of M:

$$TP\ Rate@M = correct(M)/E_{test}$$

$$FP\ Rate@M = (M - correct(M))/(0.9 \cdot E - E_{train} - correct(M)$$

We repeat the steps above for a different set of randomly removed links. The performance of a given link prediction

Table 2: Networks studied in the missing link prediction task in this paper and their properties.

network	nodes	edges	missing	density
social networks				
dolphins	62	159	16	0.084
email	1133	5452	545	0.0085
jazz	198	2742	274	0.14
connect	1095	7825	783	0.014
hep-th	8710	14254	1425	0.0003
netscience	1461	2742	274	0.0013
imdb	6260	98235	9824	0.005
technological networks				
us air	332	2126	212	0.0193
power grid	4941	6594	660	0.0004
biological networks				
protein	1870	2277	228	0.0013
c. elegans	453	2040	204	0.02

heuristic is given as Area under the ROC curve (AUC), which allows us to compare the performance of different proximity heuristics.

4.2 Data Sets

We conducted experiment on disparate undirected datasets belonging to broadly 3 categories: *Social*, *Technological* and *Biological* networks. Table 2 lists some of the statistics of the datasets.

Social networks. We studied co-authorship networks of a) theoretical physicists (**hep-th**) and b) researchers working in network science (**netscience**).[1] Each node is a scientist, and an edge between two scientists exists if they coauthor a paper together. The **hep-th** dataset was constructed on the basis of authors list of preprints available in Los Alamos e- print archive in High-Energy Theory particles area in physics between Jan 1, 1995 and December 31, 1999 [22]. The **netscience** dataset is based on preprints in the area of Network Science till May 2006 [23].

We also study the **e-mail**[2] communication network of URV University in Tarragona, Spain collected over a period of 3 months. This dataset contains around 1700 users varying from faculty, graduate students to administrators and technicians. Bulk emails which were sent to more than 50 recipients were excluded, as well as unidirectional exchanges, e.g., if A has sent a mail to B but B hasn't sent a message back to A or vice versa. An edge between two people signifies at least one email message was exchanged between the nodes from both sides [9].

The **jazz**[3] dataset contains a network of Jazz bands performing from 1912 to 1940. This dataset was obtained from Red Hot Jazz Archive.[4] The database lists the name of the musicians who performed in each band at least once in that period. An edge is created between the two Jazz bands if they have at least one musician common in both the bands [6].

The **dolphin**[5] dataset contains the social network of Bottlenose dolphins breeding in Doubtful Sound, New Zealand. This data was collected for the research program of the University of Otago-Marine Mammal Research Group. Network is compiled from the frequent and statistically significant social associations observed amongst dolphins [21].

The movie actor network, **imdb**[6], was obtained from Internet Movie Database website, which contains records of movies and their actors since 1890s. We have built a partial network of first 1000 movies where actors who have worked in the same movie are connected to each other.

The **connect** dataset represents face-to-face interactions between attendees of the 2009 Grace Hopper Conference. Each participant of the conference was given a badge. When they entered a conversation, the participants' badges were read, establishing an edge between them.[7]

Technological networks. US **power grid**[8] dataset contains a network of generators, transformers and substations in the western US power grid, which are physically connected by high-voltage transmission lines [29]. The second technological network we study is **us air**,[9] which contains airports connected by flights.

Biological networks. We also investigated biological networks, including the neural network of **c. Elegans** nematode,[10] which contains neurons that are physically bound by either a synapse or gap junction [29]. Another one is the **protein** network of yeast, which represents interactions between proteins observed in complex molecular interactions from data obtained from two-hybrid analysis [12].

4.3 Results

In this section we report the performance of link prediction heuristics on the missing link prediction task in these networks. Since all the networks studied here are undirected, some of the heuristics are mathematically equivalent: $CS = lNC$, $RA = hCS = hlNC$, and $CN = NC = hNC$.

Missing link prediction task is easiest in *social networks*, as shown in Figure 2, with AUC values approaching 1.0 in some cases. Jaccard score JC performs better than other heuristics only for *jazz* and *dolphins* datasets. In other networks, Adamic-Adar AA and resource allocation heuristic RA produce best results, because they correctly identify a higher fraction of missing links for the same number of predictions than other heuristics. Both random walk-based measure CS (and lNC) and its limited-bandwidth version

[1] http://www-personal.umich.edu/~mejn/netdata/
[2] http://deim.urv.cat/~aarenas/data/welcome.htm
[3] http://deim.urv.cat/~aarenas/data/welcome.htm
[4] www.redhotjazz.com
[5] http://www-personal.umich.edu/~mejn/netdata/
[6] http://www3.nd.edu/~networks/resources.htm
[7] Data provided courtesy of Tracy Champ.
[8] http://www-personal.umich.edu/~mejn/netdata/
[9] http://vlado.fmf.uni-lj.si/pub/networks/data/
[10] http://www-personal.umich.edu/~mejn/netdata/

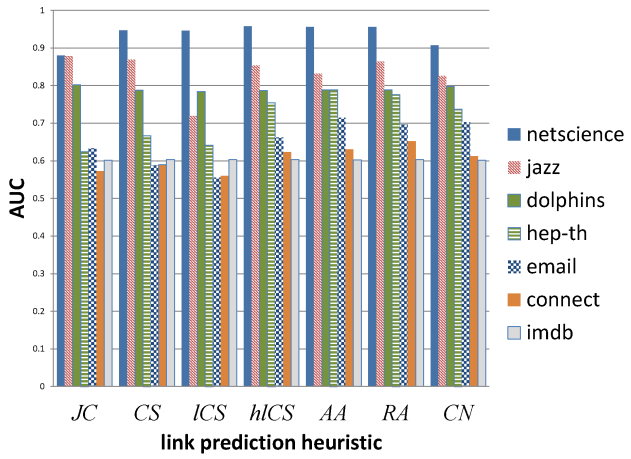

Figure 2: Performance of different link prediction heuristics on missing link task in social networks

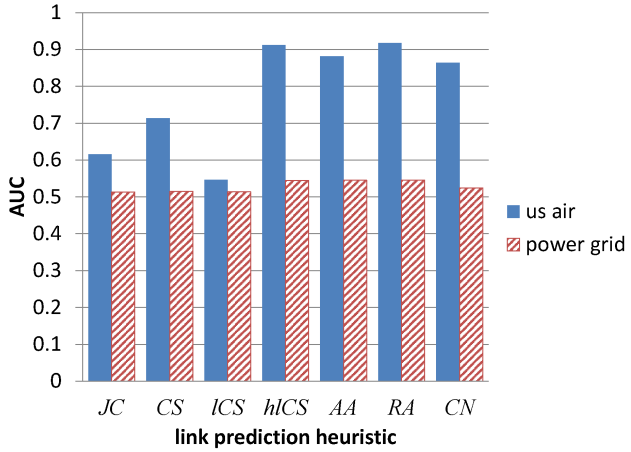

Figure 3: Performance of different link prediction heuristics on missing link task in the technological networks

lCS, along with Jaccard score (except for the *jazz* network), perform considerably worse than other heuristics.

Similar conclusions apply to the missing link prediction task in *technological networks*, shown in Figure 3. Resource allocation RA and AA heuristics produce best results, although in the *power grid* network, the common neighbors heuristics CN has better performance initially. Interestingly, links in the *us air* network are ten times easier to predict than in the *power grid* network.

In *biological networks*, shown in Figure 4, Jaccard score JC performs much worse than other heuristics, while RA and $hlCS$ lead to best performance. In the *protein* network, performance of the Adamic-Adar heuristic AA is similar to that of RA and $hlCS$.

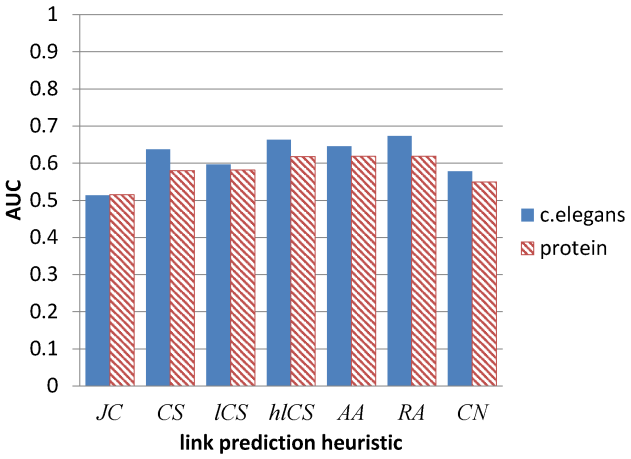

Figure 4: Performance of different link prediction heuristics on missing link task in the biological networks

4.4 Discussion of results

We compare the overall performance of different link prediction heuristics by aggregating AUC scores, Figure 5, across all datasets within each domain, giving us a sense of their relative performance. We expected epidemic-like heuristics, such as CN to work well in social networks, and random walk-based heuristics CS and RA to give better results in biological and technological networks. However, in reality, performance of different heuristics is far more nuanced. Interestingly, in social networks, there is far less difference in the performance of different link prediction heuristics than in the other domains, suggesting it is less critical which specific heuristic is used for the prediction task.

The resource allocation measure RA (and by extension hCS and $hlNC$) works consistently best across all domains, followed by AA and $hlCS$. The common neighbors heuristic CN (and also NC and hNC), which has been used widely in social network applications, does not work at all well in biological networks. Jaccard (JC), random walk-based measure CS (and lNC) and limited-bandwidth random walk (lCS) lead to the worst performance in all domains. Both RA and AA weigh the contribution of the common neighbors by the inverse of their degree (RA) or log of the degree (AA), which will tend to suppress contributions of high degree neighbors. However, penalizing high-degree source and target nodes, as is done by CS measure, does not produce good results. In fact, the limited bandwidth measure lCS, which further penalizes high-degree source and target nodes, produces even worst results. Surprisingly, its hybrid version $hlCS$, which penalizes the contribution of high-degree neighbors by the square of their degree, gives results that are competitive with RA and AA, in fact, it beats AA in biological networks.

5. RELATED WORK

In social networks, network proximity can be literally interpreted as social closeness. In his seminal paper Granovetter [8] argued that social proximity, or tie strength, which

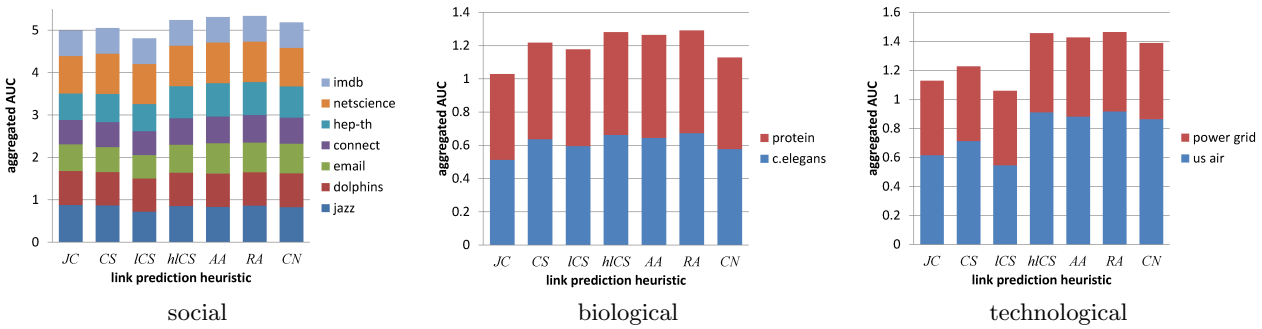

Figure 5: Aggregated performance of different link prediction heuristics.

can be defined as the intensity and the depth of interaction between two people, can be estimated from their local network structure. He proposed the number of common neighbors as the empirical measure of tie strength. Subsequently, a study of a massive mobile phone network established a correlation between social tie strength and network proximity [24]. This study measured tie strength by the frequency and duration of phone calls between two people, and it measured proximity by the fraction of common neighbors (JC). In this paper we extend Granovetter's notion of proximity to define how easily two nodes can interact of influence each other, which we argue also depends on the dynamics of interaction between nodes.

Several researchers have studied the link prediction task, in which they used network proximity to identify unobserved or missing links or to predict future links in a network. These studies used a number of measures, including the number and fraction of common neighbors, Adamic-Adar score [17, 20], as well as a measure based on resource allocation (RA) [31], and those based on the random walk, such as effective conductance [14], Local Random Walk [18], Supervised Random Walk [2] and escape probability [27, 28]. Although some measures were shown to perform better than others, no explanation was given for these differences.

There have been another set of recent studies which have used topological or attribute features for link prediction. Huang [11] proposed to look at clustering coefficient for predicting links while Koren et al. [14] used "cycle-free effective conductance" measure for improved link prediction. Zheleva et al. [30] used community information of friendship ties and family circle information to assist in predicting links. Gong et al. [7] also looked at network structure and node attribute for improving the performance of both link prediction and attribute inference problem. Another interesting work by Tang et al. [26] leveraged information from one network of coauthorship with known advisor-advisee relationship to improve prediction in a similar enterprise email network. Although these studies showed encouraging results but they are not generalizable for each network and may require certain domain knowledge to select the required attributes or information from the network.

On the link prediction task in the co-authorship networks, for example, Adamic-Adar score gave best results [17], while on the missing link prediction task in power grid and transportation networks, the linear version (RA) of Adamic-Adar performed best [31]. Sarkar et al. [25] explained why CN and AA heuristics work so well. They envision nodes residing in some latent space, where more similar nodes are closer to one other than dissimilar nodes, and prove that CN and AA estimate distance between nodes in the latent space. In this paper we argue that another dimension that has to be considered in the link prediction task is the nature of interactions between nodes. Each type of interaction, whether mediated by random walks or epidemics, leads to a different notion of network proximity, and therefore, a different link prediction heuristic. We mathematically defined heuristics based on different types of interactions, related them to known link prediction heuristics, and evaluated their performance on a variety of real-world networks. Like [20], we find that RA measure performs best on many datasets and relate it to hybrid conservative flows in networks.

6. CONCLUSION

In this paper we studied the link prediction task from the novel perspective of network flows. The flow, which mediates interaction between network nodes, determines how easily two nodes can communicate with or influence each other. We argued that network flows are linked to notions of network proximity — the closer two nodes are in a network, the easier they can interact or influence each other — with different flows leading to different notions of network proximity. We mathematically specified network proximity measures associated with different types of flows, including random walks, epidemics, limited-bandwidth walks and epidemics, and hybrid flows that are a combination of other flows, and related these to popular link prediction heuristics. Our work unifies different link prediction heuristics by viewing them as instances of network proximity under different network flows. We used these heuristics to predict missing links in a variety of social, technological and biological networks. Large variance in performance of different heuristics, especially in biological and technological networks, suggests that care should be taken in selecting an appropriate measure for each network. While the newly defined heuristics measures did not beat existing ones in the missing link prediction task, our work motivates these heuristics in terms of a flow-based framework. We postulate that the difference in performance of different heuristics lies in how well each one captures the type of flow taking place on the network.

7. REFERENCES

[1] L. Adamic and E. Adar. Friends and neighbors on the Web. *Social Networks*, 25(3):211–230, July 2003.

[2] L. Backstrom and J. Leskovec. Supervised random walks: Predicting and recommending links in social networks. *CoRR*, abs/1011.4071, 2010.

[3] L. Backstrom, E. Sun, and C. Marlow. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 61–70, New York, NY, USA, 2010. ACM.

[4] S. Borgatti. Centrality and network flow. *Social Networks*, 27(1):55–71, January 2005.

[5] S. Borgatti and M. Everett. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484, Oct. 2006.

[6] P. Gleiser and L. Danon. Community structure in jazz. *Advances in Complex Systems*, 6(4):565–573, 2003.

[7] N. Z. Gong, A. Talwalkar, L. W. Mackey, L. Huang, E. C. R. Shin, E. Stefanov, E. Shi, and D. Song. Predicting links and inferring attributes using a social-attribute network (san). *CoRR*, abs/1112.3265, 2011.

[8] M. S. Granovetter. The Strength of Weak Ties. *American Journal of Sociology*, 78(6):1360–1380, 1973.

[9] R. Guimerà, L. Danon, A. Díaz-Guilera, F. Giralt, and A. Arenas. Self-similar community structure in a network of human interactions. *Physical Review E*, 68(6):065103, 2003.

[10] N. Hodas and K. Lerman. How limited visibility and divided attention constrain social contagion. In *ASE/IEEE International Conference on Social Computing*, 2012.

[11] Z. Huang. Link prediction based on graph topology: The predictive value of generalized clustering coefficient. *Social Science Research Network Working Paper Series*, July 2010.

[12] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833):41–42, May 2001.

[13] D. Kahneman. *Attention and Effort (Experimental Psychology)*. Prentice Hall, 1973.

[14] Y. Koren, S. C. North, and C. Volinsky. Measuring and extracting proximity graphs in networks. *ACM Trans. Knowl. Discov. Data*, 1(3), Dec. 2007.

[15] R. Lambiotte, R. Sinatra, J. C. Delvenne, T. S. Evans, M. Barahona, and V. Latora. Flow graphs: Interweaving dynamics and structure. *Physical Review E*, 84(1):017102+, July 2011.

[16] K. Lerman and R. Ghosh. Network structure, topology and dynamics in generalized models of synchronization. *Physical Review E*, 86(026108), 2012.

[17] D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci.*, 58(7):1019–1031, 2007.

[18] W. Liu and L. LÃij. Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89(5):58007, 2010.

[19] W. Liu and L. Lu. Link prediction based on local random walk. *EPL (Europhysics Letters)*, 89(5):58007+, Jan. 2010.

[20] L. Lü and T. Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, Dec. 2010.

[21] D. Lusseau, K. Schneider, O. Boisseau, P. Haase, E. Slooten, and S. Dawson. The bottlenose dolphin community of doubtful sound features a large proportion of long-lasting associations. *Behavioral Ecology and Sociobiology*, 54:396–405, 2003.

[22] M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[23] M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, Sep 2006.

[24] J. P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész, and A. L. Barabási. Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18):7332–7336, May 2007.

[25] P. Sarkar, D. Chakrabarti, and A. W. Moore. Theoretical justification of popular link prediction heuristics. In *Proceedings of the Twenty-Second international joint conference on Artificial Intelligence - Volume Volume Three*, IJCAI'11, pages 2722–2727. AAAI Press, 2011.

[26] J. Tang, T. Lou, and J. Kleinberg. Inferring social ties across heterogenous networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 743–752, New York, NY, USA, 2012. ACM.

[27] H. Tong, C. Faloutsos, and Y. Koren. Fast direction-aware proximity for graph mining. In *KDD '07: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 747–756, New York, NY, USA, 2007. ACM.

[28] H. Tong, S. Papadimitriou, P. S. Yu, and C. Faloutsos. Proximity Tracking on Time-Evolving Bipartite Graphs. In *SIAM Conference on Data Mining (SDM08)*, 2008.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442, June 1998.

[30] E. Zheleva, L. Getoor, J. Golbeck, and U. Kuter. Using friendship ties and family circles for link prediction. In L. Giles, M. Smith, J. Yen, and H. Zhang, editors, *Advances in Social Network Mining and Analysis*, volume 5498 of *Lecture Notes in Computer Science*, pages 97–113. Springer Berlin Heidelberg, 2010.

[31] T. Zhou, L. Lü, and Y.-C. Zhang. Predicting missing links via local information. *The European Physical Journal B - Condensed Matter and Complex Systems*, 71(4):623–630, Oct. 2009.