Theoretical Justification of Popular Link Prediction Heuristics

Purnamrita Sarkar
Carnegie Mellon University
psarkar@cs.cmu.edu

Deepayan Chakrabarti
Yahoo! Research
deepay@yahoo-inc.com

Andrew W. Moore
Google, Pittsburgh
awm@google.com

Abstract

There are common intuitions about how social graphs are generated (for example, it is common to talk informally about *nearby* nodes sharing a link). There are also common heuristics for predicting whether two currently unlinked nodes in a graph should be linked (e.g. for suggesting friends in an online social network or movies to customers in a recommendation network). This paper provides what we believe to be the first formal connection between these intuitions and these heuristics. We look at a familiar class of graph generation models in which nodes are associated with locations in a latent metric space and connections are more likely between closer nodes. We also look at popular link-prediction heuristics such as number-of-common-neighbors and its weighted variants [Adamic and Adar, 2003] which have proved successful in predicting missing links, but are not direct derivatives of latent space graph models. We provide theoretical justifications for the success of some measures as compared to others, as reported in previous empirical studies. In particular we present a sequence of formal results that show bounds related to the role that a node's degree plays in its usefulness for link prediction, the relative importance of short paths versus long paths, and the effects of increasing non-determinism in the link generation process on link prediction quality. Our results can be generalized to any model as long as the latent space assumption holds.

1 Introduction

Link prediction is a key problem in graph mining. It underlies recommendation systems (e.g., movie recommendations in Netflix, music recommendation engines like `last.fm`), friend-suggestions in social networks, market analysis, and so on. As such, it has attracted a lot of attention in recent years, and several heuristics for link prediction have been proposed [Adamic and Adar, 2003]. In-depth empirical studies comparing these heuristics have also been conducted [Liben-Nowell and Kleinberg, 2003] and [Brand, 2005], and three observations are made consistently: (1) a simple heuristic, viz., predicting links between pairs of nodes with the most common neighbors, often outperforms more complicated heuristics, (2) a variant of this heuristic that weights common neighbors using a carefully chosen function of their degrees [Adamic and Adar, 2003] performs even better on many graphs, and (3) heuristics which weight short paths exponentially more than long paths between two nodes [Katz, 1953] often perform better than those which are more sensitive to longer paths. However, there has been little theoretical work on why this should be so. In this paper, we present a theoretical analysis of link prediction on graphs. We show how various heuristics compare against each other, and under what conditions would one heuristic be expected to outperform another. Crucially, we are able to provide theoretical justifications for all of the empirical observations mentioned above.

We define the link prediction problem as follows. There is a latent space in which the nodes reside, and links are formed based on the (unknown) distances between nodes in this latent space. Individual differences between nodes can also be modeled with extra parameters. The quality of link prediction now depends on the quality of estimation of distance between points. We show how different estimators provide *bounds* on distance. Clearly, the tighter the bounds, the better we can distinguish between pairs of nodes, and thus the better the quality of link prediction.

While any latent space model can be used, we extend a model by [Raftery *et al.*, 2002] due to two characteristics: (1) it is simple to state and analyze, (2) yet, it is powerful enough to show all of the effects that affect estimation, such as node degree, lengths of paths, etc. Our results do not assume any degree distribution on the graph; in fact, they depend on very simple properties that should be generalizable to other models as well.

Our primary contributions are as follows:

(1) We formulate the link prediction problem as a problem of estimating distances between pairs of nodes, where the nodes lie at unknown positions in some latent space and the observed presence or absence of links between nodes provides clues about their distances.

(2) We show that the number of common neighbors between a pair of nodes gives bounds on the distance between them, with the upper bound on distance decreasing quickly as the count of common neighbors increases. This justifies the popular heuristic of predicting links simply by picking node

pairs with the maximum number of common neighbors.

(3) Empirical studies [Liben-Nowell and Kleinberg, 2003] have shown that another popular heuristic [Adamic and Adar, 2003] that uses a carefully *weighted* count of common neighbors often outperforms the unweighted count. We present theoretical justification for this, and find an optimal weighting scheme (under certain assumptions).

(4) Another set of heuristics considers longer paths between pairs of nodes, e.g., hitting-time and other measures based on random walks. We show that while the number of long paths can, indeed, provide bounds on distance, these are looser than the bounds obtained if enough short paths (or ideally, common neighbors) exist. Thus, longer paths are more useful if shorter paths are rare or non-existent. The bounds obtained from long paths can be tightened given just the knowledge of *existence* of a short path.

(5) Finally, our results can be applied to any social network model where the nodes are distributed independently in some latent metric space, the probability of a link satisfies *homophily*, and links are independent of each other given node positions in the latent space.

This paper is organized as follows. In section 2 we discuss previous work on latent space models. Sections 3 and 4 prove the utility of popular heuristics like common neighbors under our latent space model for graph generation. In section 5, we analyze the implication of paths of length $\ell > 2$. Section 6 shows how to extend the analysis to handle non-determinism in the link generation process. In section 7, we summarize this paper and discuss several implications of our work.

2 The Latent Space Model

Research in sociology suggests that an important factor underlying many social networks is the notion of homophily: two nodes are more likely to have a link if they share similar characteristics [McPherson *et al.*, 2001; Faust, 1988]. These characteristics can be thought of as different features of a node, i.e. geographic location, college/university, work place, hobbies/interests etc. [Raftery *et al.*, 2002] modeled this by explicitly associating every node with a location in a D-dimensional space; links are more likely if the entities are close in latent space. All the pairwise events are independent, conditioned on their latent positions, i.e. distances in the latent space. Formally, $P(i \sim j|d_{ij}) = 1/(1 + e^{\alpha(d_{ij}-1)})$. We alter this model to incorporate a radius r in the exponent:

$$\text{(modified) RHH model:} \quad P(i \sim j|d_{ij}) = \frac{1}{1 + e^{\alpha(d_{ij}-r)}}$$

The radius r can be interpreted as the sociability of a node. Parameter $\alpha \geq 0$ controls the sharpness of the function whereas r determines the threshold. Setting $\alpha = \infty$ yields a simple deterministic model with links being formed iff nodes are less than r apart (hence, we call the finite α case the *non-deterministic* RHH model). Note, however, that given the distances the links are deterministic, but given the links, inferring the distances is an interesting problem since the node *positions* in latent space can still be stochastic. We build most of our analysis in this deterministic setting, and in section 6, we show how this analysis can be carried over to the non-deterministic case. We also extend the model to allow nodes to have distinct radii, thereby generating directed graphs; this is studied in section 4.

We assume that the nodes are uniformly distributed in a D dimensional Euclidean space. Hence $P(d_{ij} \leq x) = V(1)x^D$, where $V(1)$ is the volume of a *unit radius hypersphere*. This uniformity assumption has been made in earlier social network models, e.g. by [Kleinberg, 2000], where the points are assumed to lie on a two dimensional grid. In order to normalize the probabilities, we assume that all points lie inside a *unit volume hypersphere* in D dimensions. The maximum r satisfies $V(r) = V(1)r^D = 1$.

Connection to the Link Prediction Problem. A latent space model is well-fitted for link prediction because, for a given node i, the most likely node it would connect to is the node closest to it in latent space (that it is not already linked to). Thus, the predicting distances between a pair of nodes is the key. While this can be obtained by maximizing the likelihood of the underlying statistical model, we show that one can obtain high probability bounds on distances from graph based heuristics. In fact we show that the distance to the node picked using a popular heuristic is within a small factor of the *true distance*. This factor quickly goes to zero as N becomes large. Although our analysis uses an extension of the RHH model to actually obtain the bounds on distances, the only property we use is of homophily in a latent metric space, i.e. if two nodes are *close* in some social space, then they are likely to form a link. Hence this idea should carry over to other social network models as well.

3 Deterministic Model with Identical Radii

Consider a simple version of the RHH model where all radii are equal to r, and $\alpha \to \infty$. This implies that two nodes i and j share a link (henceforth, $i \sim j$) iff the distance d_{ij} between them is constrained by $d_{ij} < r$. Thus, given node positions, links are deterministic; however the node positions are still non-deterministic. While this might appear to be a strong constraint, we will show later in Section 6 that similar results are applicable even for finite but large α. We now analyze the simplest of heuristics: counting the common neighbors of i and j. Let there be N nodes in total.

Let $\mathcal{N}(i)$ be the set of neighbors of node i. Let Y_k be a random variable which is 1 if $k \in \mathcal{N}(i) \cap \mathcal{N}(j)$, and 0 otherwise. Given d_{ij}, $\forall k \notin \{i, j\}$, the Y_k's are independent since they only depend on the position of point k. This gives:

$$E[Y_k|d_{ij}] = P(i \sim k \sim j|d_{ij})$$
$$= \int_{d_{ik},d_{jk}} P(i \sim k|d_{ik})P(j \sim k|d_{jk})P(d_{ik}, d_{jk}|d_{ij})d(d_{ik})d(d_{jk})$$

In the deterministic model, this quantity is exactly equal to the volume of intersection of two balls of radius r centered at i and j (see Figure 1). Denote this volume by $A(r, r, d_{ij})$. Also, the observed value of $\sum_k Y_k$ is simply the number of common neighbors η. From now on we will drop the d_{ij} part when we write expectation for notational convenience. However any expectation in terms of area of intersection is obviously computed *given the pairwise distance* d_{ij}. Thus by

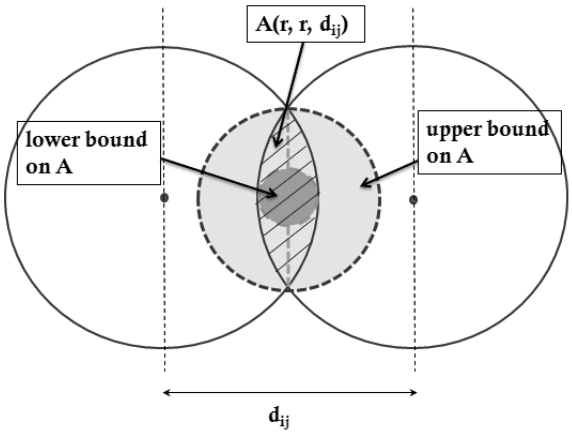

Figure 1: Common neighbors of two nodes must lie in the intersection $A(r, r, d_{ij})$.

using empirical Bernstein bounds [Maurer and Pontil, 2009], we have:

$$P\left[\frac{\eta}{N} - \epsilon \leq A(r, r, d_{ij}) \leq \frac{\eta}{N} + \epsilon\right] \geq 1 - 2\delta, \quad (1)$$

where $\epsilon = \sqrt{2\mathtt{var}_N(Y)\log 2/(\delta N)} + 7\log 2/(3\delta(N-1))$, and $\mathtt{var}_N(Y)$ is the sample variance of Y. Now, $A(r, r, d_{ij})$ is a monotonically decreasing function of d_{ij}, so Equation 1 can be used to infer bounds on d_{ij}. One such bound, which we state without proof, is as follows.

$$2r\left(1 - \left(\frac{\eta/N + \epsilon}{V(r)}\right)^{1/D}\right) \leq d_{ij} \leq 2r\sqrt{1 - \left(\frac{\eta/N - \epsilon}{V(r)}\right)^{2/D}}$$

Using common neighbors in link prediction. Recall that in link prediction, we want to pick the node which is most likely to be a neighbor of i, and is not currently a neighbor (call this OPT). If we knew all node positions, we would pick the non-neighbor with the minimum distance (d_{OPT}). However, since positions in latent space are unknown, we instead predict a link to the node that shares the most common neighbors with i (call this MAX). We show next that the distance to the node with largest common neighbors (d_{MAX}) is within an additive factor of d_{OPT}, and this factor goes to zero as N increases. This implies that, as N increases, link prediction using the number of common neighbors converges to the optimal prediction.

Let the number of common neighbors between i and OPT be η_{OPT}, and between i and MAX be η_{MAX}. The corresponding distances are d_{OPT} with d_{MAX}, and the corresponding areas of intersection with i are A_{OPT} and A_{MAX} respectively. Let $\epsilon_o = \sqrt{2\mathtt{var}_N(Y_{OPT})\log 2/(\delta N)} + 7\log 2/(3\delta(N-1))$ and $\epsilon_m = \sqrt{2\mathtt{var}_N(Y_{MAX})\log 2/(\delta N)} + 7\log 2/(3\delta(N-1))$, where Y_{OPT} and Y_{MAX} denote the random variable for common neighbors between i and OPT, and i and MAX respectively. Let $\epsilon_f = \epsilon_o + \epsilon_m$.

Theorem 3.1.

$$d_{OPT} \leq d_{MAX} \overset{w.h.p}{\leq} d_{OPT} + 2r\left(\frac{\epsilon_f}{V(r)}\right)^{\frac{1}{D}} \leq d_{OPT} + 2\left(\frac{\epsilon_f}{V(1)}\right)^{\frac{1}{D}}$$

As $N \to \infty$, $\epsilon_f \to 0$, so the node with the highest number of common neighbors will be the optimal node for link prediction.

4 Deterministic Model with Distinct Radii

Until now our model has used the same r for all nodes. The degree of a node is distributed as $Bin(N, V(r))$, where $V(r)$ is the volume of a radius r. Thus r determines the degree of a node in the graph, and identical r will lead to a roughly regular graph. In practice, social networks are far from regular. In order to accommodate complex networks we will now allow a different radius (r_i) for node i. For this section, we will assume that these radii are independent, and are known. The new connectivity model is: $i \to j$ iff $d_{ij} \leq r_j$, where $i \to j$ now represents a *directed* edge from i to j. While variants of this are possible, this is similar in spirit to a citation network, where a paper i tends to cite a well-cited paper j (with larger number of in-neighbors) than another infrequently cited paper on the same topic; here, r_j can be thought of as the measure of popularity of paper j. Under this model, we will show why some link prediction heuristics work better than others.

As in the previous section, we can use common neighbors to estimate distance between nodes. We can count common neighbors in 4 different ways as follows:

- **Type-1:** All k, s.t. $k \to i$ and $k \to j$: all nodes which point to both i and j. The probability of this given d_{ij} is $P(d_{ik} \leq r_i \cap d_{jk} \leq r_j | d_{ij})$, which can be easily shown to be $A(r_i, r_j, d_{ij})$.
- **Type-2:** All k, s.t. $i \to k$ and $j \to k$: all nodes to which both i and j point. The probability of this given d_{ij} is $A(r_k, r_k, d_{ij})$.
- **Type-3:** All k, s.t. $i \to k$ and $k \to j$: all directed paths of length 2 from i to j. The probability of this given d_{ij} is given by $A(r_k, r_j, d_{ij})$.
- **Type-4:** All k, s.t. $j \to k$ and $k \to i$: all directed paths of length 2 from j to i. The probability of this given d_{ij} is given by $A(r_i, r_k, d_{ij})$.

If we count type-1 nearest neighbors, the argument from section 3 carries over, and if there are enough common neighbors of this type, we can estimate d_{ij} by computing $A(r_i, r_j, d_{ij})$. However, if both r_i and r_j are small, there might not be many common neighbors; indeed, if $d_{ij} > r_i + r_j$, then there will be no type-1 common neighbors. In such cases, we consider type-2 neighbors, i.e. the ones which both i and j point to. The analysis for type-3 and type-4 neighbors is very similar to that for type-2, and hence we do not discuss these any further. In the type-2 case, the radii r_k of the common neighbors play an important role. Intuitively, if both i and j point to a very popular node (high radius r_k), then that should not give us a lot of information about d_{ij}, since it is not very surprising. In particular, any type-2 common neighbor k leads to the following constraint: $d_{ij} \leq d_{ik} + d_{jk} \leq 2r_k$. Obviously, the bound is stronger for small values of r_k. This argues for

weighting common neighbors differently, depending on their radii. We formalize this intuition using a toy example below, while noting that the analysis in the following section can be generalized to graphs where the radii of the nodes form a finite set.

Motivating Example. Take a toy network where the nodes can have two different radii R and R', with $R < R'$. The total number of low radii nodes is N_R, whereas that of large radii nodes is $N_{R'}$.

The formula for the expectation of the number of type-2 common neighbors will now have a mixture of $A(R, R, d_{ij})$ and $A(R', R', d_{ij})$. One solution is to estimate high probability bounds on distances from the two different classes of common neighbors separately, and then examine the intersection of these bounds. Another solution is to look at weighted combinations of common neighbors from different radii. The weights will reflect how important one common neighbor is relative to another. For example, consider a pair of papers which both cite a book on introduction to algorithms (cited by 5000 other papers, i.e. higher radius), and a specific article on randomized algorithms (cited by 30 other papers, i.e. lower radius). The second article gives more evidence on the "closeness" or similarity of the pair. We will formalize this approach next.

Suppose we observe η_R common neighbors of N_R nodes of small radius, and $\eta_{R'}$ common neighbors of $N_{R'}$ nodes of large radius, between pair of nodes i, j. The likelihood of these observations, given the pairwise distance d_{ij} is:

$$\prod_{r \in \{R, R'\}} \binom{N_r}{\eta_r} A(r, r, d_{ij})^{\eta_r} (1 - A(r, r, d_{ij}))^{N_r - \eta_r} \quad (2)$$

We want to rank pairs of nodes using the distance estimate d^* that maximizes the likelihood of this partial set of observations. However, if $\eta_R > 0$, the logarithm of the above is defined only when $d_{ij} \leq 2R$. To make the likelihood well-behaved, we introduce a small noise parameter β: node i connects to node j with probability $1 - \beta$ (if $d_{ij} \leq r_j$), or with probability β (otherwise). Now, the probability of having a type-2 common neighbor of radius r will be $\beta + A(r, r, d_{ij})(1 - \beta)$. For ease of exposition we will denote this by $A_\beta(r, r, d_{ij})$. The new likelihood will be exactly as in eq. (2), except we will use A_β instead of A. Setting the derivative of the logarithm to zero yields:

$$w(R, d^*) N_R A_\beta(R, R, d^*) + w(R', d^*) N_{R'} A_\beta(R', R', d^*)$$
$$= w(R, d^*) \eta_R + w(R', d^*) \eta_{R'}$$

where, $w(R, d^*) = \dfrac{-\left.\frac{dA_\beta(R,R,d_{ij})}{d_{ij}}\right|_{d^*}}{A_\beta(R, R, d^*)(1 - A_\beta(R, R, d^*))}$. Note that the negative sign is only to make both sides positive, since A_β decreases with distance.

Further analysis of this equation ([Sarkar *et al.*, 2010]) shows that for a given distance d and increasing radius r, the weight $w(r, d)$ first decreases sharply but increases again once r becomes close to the maximum radius, i.e., $V(r) \approx 1$. Thus, it is high for both nodes of very low and very high radius. The reason for giving high weight to a low-radius common neighbor is clear: the presence of such a neighbor gives strong evidence that d is small. On the other hand, the *absence* of a *very* high degree node in the set of common neighbors gives strong evidence that d is very large, which is why $w(r, d)$ is large for common neighbors of very high degree. Note that the presence of low radius common neighbors in the absence of very high radius common neighbors is extremely unlikely. This is because if a pair of nodes are close enough to connect to a low radius node, they are also very likely to both be within the radius of some very high radius node.

Since $NV(r)$ is the expectation of the indegree of a node of radius r, high-radius nodes are expected to have extremely high degrees. However, high-degree nodes in real-world settings typically connect to no more than $10 - 20\%$ of the set of nodes, which is why a practical weighting only needs to focus on situations where $V(r) \ll 1$. For relatively small d (which are the interesting candidates for link prediction), the weights $w(r, d)$ can then well approximated by $w(r) = 1/r$ up to a constant. Note that this is identical to weighting a node by $1/(NV(r))^{1/D}$, i.e., essentially weighting a common neighbor i by $1/\deg(i)^{1/D}$.

The Adamic/Adar link prediction heuristic. The Adamic/Adar measure [Adamic and Adar, 2003] was introduced to measure pairwise web-page similarity. Instead of computing the number of common features of two webpages, the heavier weights were assigned to the rarer features. In our social networks context, this translates to:

$$\texttt{Adamic/Adar} = \sum_{k \in \mathcal{N}(i) \cap \mathcal{N}(j)} \frac{1}{\log(\deg(k))}$$

Note that this has the same trend as the $1/\deg(i)^{1/D}$ formula discussed above. [Liben-Nowell and Kleinberg, 2003] have shown that Adamic/Adar out-performs the number of common neighbors in terms of link prediction accuracy in a variety of social and citation networks, confirming the positive effect of a skewed weighting scheme that we observed in the motivating example.

We can analyze the Adamic/Adar measure as follows. In our model, the expected degree of a node k of radius r_k is simply $NV(r_k)$, so we set the weights as $w_k = 1/\log(NV(r_k))$. Let $\mathcal{S} = \sum_k w_k Y_k$, where random variable $Y_k = 1$ if k is a type-2 common neighbor of i and j, and zero otherwise. Clearly, $E[\mathcal{S}] = \sum_k w_k A(r_k, r_k, d_{ij}) = \sum_k A(r_k, r_k, d_{ij})/\log(NV(r_k))$. Let the minimum and maximum radii be $r_{\min}$ and $r_{\max}$ respectively. The following can be easily obtained from the Chernoff bound:

Lemma 4.1. $\frac{\mathcal{S}}{N}\left(1 - \sqrt{\frac{3\log(NV(r_{\max}))\ln(1/\delta)}{N \cdot A(r_{\max}, r_{\max}, d_{ij})}}\right) \leq \frac{E[\mathcal{S}]}{N} \leq \frac{\mathcal{S}}{N}\left(1 + \sqrt{\frac{3\log(NV(r_{\min}))\ln(1/\delta)}{N \cdot A(r_{\min}, r_{\min}, d_{ij})}}\right)$

Clearly, the error terms decay with increasing N, and for large N, we can tightly bound $E[\mathcal{S}]$. Since $E[\mathcal{S}]$ is monotonically decreasing function of d_{ij}, this translates into bounds on d_{ij} as well.

We have seen that low radius common neighbors imply that distance is small, whereas fewer high degree common neighbors in the absence of any low degree common neighbors imply that distance is large. Based on these observations we can

also examine the number of common neighbors, say Q_R (T_R) with radii smaller (larger) than a given radius R. This translates into bounds on distance, the key idea being that large Q_R for small R indicates a small distance, whereas small T_R for large R indicates large distance with high probability. In addition, we design ([Sarkar *et al.*, 2010]) *Sweep Estimators* that obtain bounds on distance from the entire range of radii of common neighbors and retains the best bounds.

5 Estimators using Longer Paths in the Deterministic Model

The bounds on the distance d_{ij} described in the previous sections apply only when i and j have common neighbors. However, there will be no common neighbors if (for the undirected case) (a) $d_{ij} > 2r$, or (b) no points fall in the intersection area $A(r, r, d_{ij})$ due to small sample size N. In such cases, looking at paths of length $\ell > 2$ between i and j can yield bounds on d_{ij}. Even if common neighbors exist, these bounds can be tighter, especially when there are few common neighbors but many longer paths.

An Upper Bound for $\ell > 2$. Recall that a *simple* path of length ℓ from i to j is a path of the form $i \sim k_1 \sim k_2 \sim \ldots k_{\ell-2} \sim j$, with no repeated node. We need to infer bounds on d_{ij} given the observed number of simple paths $\eta_\ell(i, j)$.

As before, we can derive these by bounding the expected number of ℓ-hop paths $\eta_\ell(i, j)$, and the deviation of the true number of paths from its expectation. However, for $\ell > 2$, the paths are dependent (e.g., two paths can share intermediate nodes) so Chernoff bounds do not apply. Instead, we proceed in three steps. First, we bound the maximum degree Δ of any graph generated by the RHH model, and show that $\eta_\ell(i, j) \leq \Delta^{\ell-1}$. Second, we upper-bound $E[\eta_\ell(i, j)]$ by *triangulation*. As shown in Figure 2, a given sequence of distinct points $(i, k_1, k_2, \ldots, j)$ forms a path if each k_i is within distance $a_i \leq r$ of the previous point k_{i-1} and within distance d_i of j, and the intermediate distances d_i can be bounded in terms of r and d_{ij} by repeated applications of the triangle inequality. The bounds on a_i and d_i can, in turn, be translated into bounds on the probability of the given sequence forming a path in the graph. Since there at most $\Delta^{\ell-1}$ such paths, this gives a bound on $E[\eta_\ell(i, j)]$. Finally, we observe that $\eta_\ell(i, j)$ is "robust" in the sense that changing the position of any one point in the latent space can cause only a bounded change in $\eta_\ell(i, j)$. We utilize this to bound the maximum deviation of $\eta_\ell(i, j)$ from its expectation. Combining these, we obtain the following result.

Theorem 5.1. *With probability at least* $(1 - 2\delta)$,

$$\eta_\ell(i,j) \leq (NV)^{\ell-1} \left[\prod_{p=1}^{\ell-1} A(r, p \times r, (d_{ij} - (\ell - p - 1)r)_+) + \frac{(\ell-1)\sqrt{\frac{1/\delta}{2}}}{\sqrt{NV}\left(1 + \sqrt{\frac{\ln(N/\delta)}{2NV}}\right)} \right] \left(1 + \sqrt{\frac{\ln(N/\delta)}{2NV}}\right)^{\ell-1},$$

where $x_+ = \max(x, 0)$.

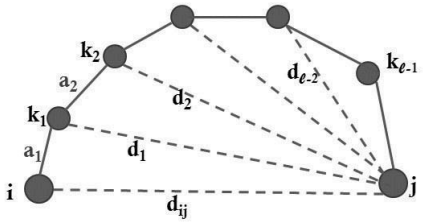

Figure 2: Triangulation for bounding d_{ij} using ℓ-hop paths.

Bounding d_{ij}. Theorem 5.1 yields an upper bound d_{ij} as follows. Only the first term in the summation depends on d_{ij}, and this term decreases monotonically with increasing d_{ij}. Thus, a simple binary search can give us the value of d_{ij} that achieves the equality in Theorem 5.1, and this is an upper bound on d_{ij}.

A looser but analytic bound can be obtained by upper-bounding all but one of the $A(.)$ terms by 1. For example, for 3-hop paths, we have:

$$d_{ij} \leq r + 2r\sqrt{1 - \left(\frac{\eta_3(i,j)/c(N,\delta) - c'(N,\delta)}{V(r)}\right)^{2/D}}$$

In general, bounds for ℓ-hop paths are of the form $d_{ij} \leq \ell r(1 - g(\eta_\ell(i, j), \epsilon))$. Thus, for some $\ell' > \ell$, $\eta_{\ell'}(i, j)$ needs to be much larger than $\eta_\ell(i, j)$ for the bound using ℓ' to be stronger than that for ℓ. In particular, this shows that when enough common neighbors are present (i.e., 2-hop paths), looking at longer paths is unlikely to improve bounds and help link prediction, thus theoretically confirming the empirical observations of [Liben-Nowell and Kleinberg, 2003]. However, we note that tighter bounds can be found when shorter paths are known to exist; this is discussed in detail in [Sarkar *et al.*, 2010].

Observations. Our analysis of ℓ-hop paths yields the following observations. (1) When short paths are non-existent or rare, the bounds on d_{ij} that we obtain through them can be loose. Longer paths can be used to yield better bounds in such cases. (2) As ℓ increases, more and more long paths need to be observed before the corresponding bound on d_{ij} becomes comparable or better than bounds obtained via shorter paths. (3) Even the *existence* of a short path can improve upper bounds obtained by all longer paths. In addition, lower bounds on d_{ij} can also be obtained. (4) The *number* of paths is important to the bound. Link prediction using the length of the shortest path ignores this information, and hence should perform relatively poorly, as observed by [Liben-Nowell and Kleinberg, 2003; Brand, 2005] and [Sarkar and Moore, 2007].

6 The Non-deterministic Case

Previously we have assumed that, given the positions of points, the corresponding graph could be inferred exactly. In terms of the RHH model introduced in section 2, this corresponds to setting $\alpha \to \infty$. In this section, we investigate the effects of finite α. Our analysis shows that while bounds become looser, the results are still qualitatively similar.

The core idea underlying almost all of our previous results has been the computation of the probability of two nodes i and j having a common neighbor. For the deterministic case, this is simply the area of intersection of two hyperspheres, $A(r, r, d_{ij})$, when all nodes have the same radius r. However, in the non-deterministic case, this probability is hard to compute exactly. Instead, we can give the following simple bounds on $Pr_2(i, j)$, which is the probability of observing a common neighbor between two nodes i and j that are distance d_{ij} apart and have identical radius r.

Theorem 6.1.

$$Pr_2(i,j) > \tfrac{1}{4}\left(A(r,r,d_{ij}) + 2e^{-\alpha d_{ij}} \cdot (V(r) - A(r,r,d_{ij}))\right)$$

$$Pr_2(i,j) < \begin{cases} A(r,r,d_{ij}) + 2V(r) \cdot \dfrac{\left[1 - \left(\frac{D}{\alpha r}\right)^D\right]}{\frac{\alpha r}{D} - 1} & \text{(for } \alpha r > D\text{)} \\ A(r,r,d_{ij}) + 2D \cdot V(r) & \text{(for } \alpha r = D\text{)} \\ A(r,r,d_{ij}) + 2V(D/\alpha) \cdot \dfrac{\left[1 - \left(\frac{\alpha r}{D}\right)^D\right]}{1 - \frac{\alpha r}{D}} & \text{(for } \alpha r < D\text{)} \end{cases}$$

Observations and Extensions. The importance of theorem 6.1 is that the probability of observing a common neighbor is still mostly dependent on the area of intersection of two hyperspheres, i.e. $A(r, r, d_{ij})$. However, there is a gap of a factor of 4 between the lower and upper bounds. This can still be used to obtain reasonable bounds on d_{ij} when enough common neighbors are observed. However, when we consider longer paths, the gap increases and we might no longer be able to get strong bounds.

The reason for this is that theorem 6.1 only uses the fact that probability of linking i and j is at least $1/2$ when d_{ij} is less than r. This statement is applicable to all α. However, we typically want to perform link prediction only when α is large, as small values of α yield graphs that are close to random and where no link prediction methods would work. For the case of large α, we can get much stronger lower bounds and close the factor-of-4 gap, as follows.

To compute the probability $Pr_2(i, j)$, we need to integrate the product of the link probabilities over the intersection of the two hyperspheres of radius r around nodes i and j. Let this region be denoted by $S(i, j)$. Suppose that, instead of integrating over $S(i, j)$, we integrate over a smaller subset $S'(i, j)$. While the volume of $S'(i, j)$ would be smaller, the minimum probabilities inside that subset could be much higher than in $S(i, j)$, leading to a better overall lower-bound. We consider $S'(i, j) = \{x_k | d_{ik} < r', d_{jk} < r'\}$ to be the intersection of two hyperspheres of radius $r' < r$, centered on i and j.

$$\forall r' < r, Pr_2(i, j, x_k \in S'(i, j)) \geq \left(\frac{1}{1 + e^{\alpha(r'-r)}}\right)^2 \cdot \text{vol}(S'(i, j))$$

Ideally, we would like to pick r' to maximize this, but $\text{vol}(S'(i, j))$ depends on d_{ij} as well. Instead we propose the following heuristic:

$$\text{Pick } r' \text{ to maximize } \left(\frac{1}{1 + e^{\alpha(r'-r)}}\right)^2 \cdot V(r') \qquad (3)$$

Lemma 6.2. *If $\alpha > D/r$ and r' is picked according to Equation 3, then $r' < r$.*

Thus, for large enough α, we can find a good r' which can improve the gap between upper and lower bounds of $Pr_2(i, j)$. The optimal r' gets closer to r as α increases, but its exact value has to be obtained numerically.

7 Summary and Discussion

The paper presents a theoretical study of link prediction and the heuristics commonly used for that purpose. We formalize the link prediction problem as one of estimating distances between nodes in a latent space, where the observed graph structure provides evidence regarding the unobserved positions of nodes in this space. We present theoretical justifications of two common empirical observations: (1) the simple heuristic of counting common neighbors often outperforms more complicated heuristics, (2) a variant that weights common neighbors by the inverse of the logarithm of their degrees [Adamic and Adar, 2003] often performs better. We show that considering longer paths is useful only if shorter paths (especially, common neighbors) are not numerous enough for the bounds obtained from them to be tight enough. However, the bounds obtained from longer paths can be made significantly tighter if short paths are known to exist.

References

[Adamic and Adar, 2003] Lada A. Adamic and Eytan Adar. Friends and neighbors on the web. *Social Networks*, 25, 2003.

[Brand, 2005] M. Brand. A Random Walks Perspective on Maximizing Satisfaction and Profit. In *SIAM '05*, 2005.

[Faust, 1988] Katherine Faust. Comparison of methods for positional analysis: Structural and general equivalences. *Social Networks*, 1988.

[Katz, 1953] L. Katz. A new status index derived from sociometric analysis. In *Psychometrika*, 1953.

[Kleinberg, 2000] J. Kleinberg. The Small-World Phenomenon: An Algorithmic Perspective. In *STOC*, 2000.

[Liben-Nowell and Kleinberg, 2003] David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *CIKM '03*, 2003.

[Maurer and Pontil, 2009] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample-variance penalization. In *Conference on Learning Theory*, 2009.

[McPherson *et al.*, 2001] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 2001.

[Raftery *et al.*, 2002] A. E. Raftery, M. S. Handcock, and P. D. Hoff. Latent space approaches to social network analysis. *J. Amer. Stat. Assoc.*, 15:460, 2002.

[Sarkar and Moore, 2007] Purnamrita Sarkar and Andrew Moore. A tractable approach to finding closest truncated-commute-time neighbors in large graphs. In *Proc. UAI*, 2007.

[Sarkar *et al.*, 2010] Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew Moore. Theoretical justification of popular link prediction heuristics. In *COLT*, 2010.