Navigation in Real-World Complex Networks through Embedding in Latent Spaces

Xiaomeng Ban* Jie Gao* Arnout van de Rijt†

Abstract

Small-world experiments in which packages reach addressees unknown to the original sender through a forwarding chain confirm that acquaintance networks have short paths, a property that was later also discovered in many other networks. They further show that people can find these paths by passing the package on to the acquaintance most socially proximate to the target. This has led researchers to conjecture that perhaps also in many other networks some proximity-based algorithm can be used to find short paths, provided that nodes are given appropriate coordinates. Although potential applications are numerous, ranging from decentralized search to recommendation-based trust to disease control, this conjecture has remained largely unverified. In this paper we apply algorithmic methods to embed nodes in some latent space and employ greedy routing to deliver packages. Using these methods we empirically investigate the navigability of five real-world complex networks from diverse contexts and of varying topology. In each network, we deliver a majority of packages in fewer than six hops.

1 Introduction

In the 1960's, Stanley Milgram and his collaborators conducted a series of experiments in which individuals from Nebraska and Kansas were asked to try and get letters delivered to unknown recipients in Boston [40]. A person forwards the letter to a friend who is more likely to know the target. Many letters were discarded by uncooperative intermediaries, but about 20% of the letters arrived at the target, in an average of under six hops. This experiment is the earliest to verify the 'small-world phenomenon' (aka 'six degree of separation') that there *exists* a short path between almost any pair of individuals in the world. It was later discovered that many other networks, in vastly different contexts ranging from power grids, film collaboration networks, and neural networks [20] to email networks [17], food webs [42] and protein interaction networks [22], also exhibit the small-world property.

In addition to revealing the existence of short paths in real-world acquaintance networks, the small-world experiments showed that these networks are *navigable*: A short path was discovered through a *local* algorithm with the participants forwarding to a friend who they believed to be more likely to know the target. Although forwarding decision-making was not systematically recorded, geographical proximity was found to be an important forwarding criterion in some cases. Other criteria such as profession and popularity may have been used as well. A recent small-world study using email-chains [17] confirms this, finding that at least half of the choices were due to either geographical proximity of the acquaintance to the target or occupational similarity. Thus these experiments hint that perhaps also in other networks, some greedy routing algorithm can successfully deliver messages, provided that nodes are given appropriate coordinates.

In this paper we consider the conjecture that real-world networks from diverse contexts, social and non-social, can be embedded in a *low-dimensional hidden space* where the distances between nodes in the hidden space approximate their graph distances in the network, such that some greedy mechanism minimizing the distances to the destination *in the latent space* is able to find a short path for most pairs of nodes. The existence of such a latent space in *social* networks is suggested by the sociological principle of *homophily*, that friends tend to have similar traits or adopt similar behaviors [34]. A set of individuals with a large number of social ties between them may indicate that they have nearby positions in the space of characteristics [5]. This social space may refer to a space of observed or unobserved latent characteristics that represent potential transitive tendencies in network relations. Moreover, networks may have a spatial component, with connectivity decreasing in geographical distance [13, 30]. Nevertheless, we must remark that the methods we consider do not use any social, spatial or any other node-identifying information. Instead, they attempt to *discover* this space from network structure alone. This allows them to be applicable to anonymous social networks as well as non-social networks that must be navigated. We must also emphasize that these methods are rather basic and straight-forward. The methods themselves are not the main contribution here. Rather, we use them to verify the navigability conjecture on five different empirical networks, social and non-social, of very diverse origin and topology. Our results show that they can all be navigated. In each network, a majority of packages is delivered in fewer than six steps.

*Department of Computer Science, Stony Brook University. {xban,jgao}@cs.sunysb.edu.

†Department of Sociology, Stony Brook University. arnout.vanderijt@stonybrook.edu.

Our contribution. We propose to first embed the network in a latent space and apply greedy routing with the coordinates generated.

We start with the obvious choice of Euclidean space and use the prominent embedding algorithm of Multi-Dimensional Scaling (MDS) [7]. We measure the distance between any two nodes as the length of the shortest path between them in the social network. With the all pair graph distances MDS produces a coordinate for every node in the Euclidean space with a pre-specified dimensionality.

MDS however requires the all pairs distance matrix of the network and is computationally intensive. To make our method feasible for very large empirical networks that are only partially observed, we adopt landmark-based MDS (LMDS) [16], in which a few nodes are selected as landmarks and embedded first, and the rest of the nodes embed themselves using distances to these landmarks. It is shown below that landmark MDS achieves a comparable performance (in terms of delivery rate and routing path length) to that of MDS, but more than an order of magnitude faster. The use of landmark MDS also implies a distributed implementation of the embedding/routing algorithm. In particular, we can sample a small set of nodes in the social network as landmarks, and embed them. Any individuals who would like to route can embed themselves *on the fly*, by using the distances to the landmarks. In a coauthorship or film collaboration network, the obvious choices for the landmarks are the famous scholars or actors with well-known connections to each other and the distances to others pre-calculated (such as the Erdös number or the Bacon number). The distances from all nodes to the landmarks can be computed in time linear in network size.

As the Euclidean space of dimension d has a geometric growth rate and small-world graphs have been observed to have low diameters, suggesting an exponential growth rate, we also consider embedding into the hyperbolic space. We employ R. Kleinberg's embedding method [28] to embed a tree in hyperbolic space with the induced coordinates used for greedy routing.

We consider five real-world networks from diverse contexts and of varying topology: A peer-to-peer file sharing network, a scientific collaboration network, a movie-actor coappearance network, and an Internet autonomous systems network. Surprisingly, with simply out-of-the-shelf methods one is able to get reasonably high delivery rate, and even more, very small average path length, within *six* steps. Before we conducted these experiments, we expected that possibly some fraction of the messages can reach the destination via the greedy algorithm, as the embedding by MDS preserves the distances to some extent. We have never expected that these messages only use 4 or 5 hops on average! Note that successful delivery does not imply the path is short.

Significance of network navigation. The task of identifying short paths appears in a wide variety of empirical settings. Short paths allow for speedy package delivery in decentralized file-sharing networks [2], searching for pathways in very large metabolic networks [18], and enable reputation-based trust in exchange [11]. The fact that we are living in a 'small world' suggests that potentially we can consult with any expert in a field of interest, or do business with any individual, with recommendations through a short chain of friends — if only we were able to find such a short path quickly. Short path identification will also facilitate a number of social operations. The occupants 'structural holes', positioned on short paths between otherwise distant others in social networks earn brokerage benefits [10, 41]. The task of ensuring sufficient structural distance between two individuals appears frequently: monopoly mediation, double-targeting in advertisement, avoiding infection, and sharing confidential information that must travel far to reach an unwanted ear [41, 12].

In many of these application scenarios, a central navigation device is often lacking, distributed flooding causes congestion and excessive use of resources. In addition, the nodes may have limited information storage capacity and processing power. The network data may be incomplete, and node attribute information may be absent altogether. For these settings greedy routing is a better choice than centralized short path computations.

2 Related Work on Small World Graphs

2.1 Navigation in model networks A number of studies have proposed mathematical models for small world networks and navigation in such networks.

Watts and Strogatz [20] proposed as a 'random rewiring model' in which with some probability the edges on a ring are rewired to random vertices. The rewiring probability can be tuned to generate networks in between the two extremes of perfectly regular and perfectly random networks. It is shown that for most of the parameter space, networks simultaneously exhibit high clustering and low path length. Additionally, three diverse real-world networks are shown to exhibit both properties. They show that short paths often exist but not how they could be found without global knowledge of the network.

Barabasi *et al.* [3] considered an evolving graph in which each newcomer connects to existing vertices with probability proportional to their current degree (thus the name preferential attachment model). The network constructed is a scale-free graph, i.e., it has a power-law degree distribution, a property of various real-world networks. The graph also has small diameter and in addition, hub nodes that are highly connected to other vertices. For scale-free graphs, a degree-based greedy routing has been investigated [2, 24]. The intuition is to send the message to a neighbor with higher degree as the neighbor is more likely to be a neighbor of the

destination.

Another idea to navigate in small world networks is to make use of user identities (geographical location, profession, etc.) and the structure of the 'social space'. Kleinberg [27] considered a lattice network in $\mathbb{R}^d$ and placed additional edges pq with probability proportional to $1/|pq|^\alpha$, where $|pq|$ is the Euclidean distance between p, q and α is a parameter. Then he showed that if $\alpha = d$ the greedy algorithm of delivering the message to the node closest to the destination in *Euclidean distance* is able to find a short path to the destination with polylogarithmic number of hops. If $d \neq \alpha$, the greedy routing takes necessarily polynomial number of hops, i.e., the network is not navigable. Watts *et al.* [19] considered a hierarchical professional organization of individuals and a homophilious network with ties added between two nodes closer in the hierarchy with a higher probability. If each node has a fixed probability of dropping the message, they show a greedy routing algorithm sending packages to the neighbor most similar to the target (called homophily-based routing) successfully deliver a fraction of the messages before they are dropped. Kleinberg [25] also confirmed similar results on a hierarchical network. Şimşek and Jensen [39] evaluated routing schemes on networks with different homophilous level. When the homophily level is low, degree based routing is effective as the hubs connect different part of the network. When the network homophily level is high, hubs are not very useful as they connect to other individuals very similar to themselves. They proposed to use a simple product of the homophily and degree to estimate the neighbor who is most likely to be directly connected to the target.

Boguñá et al. [6] incorporated both the idea of having a social space and the power law degree distribution. They considered nodes on a ring and assigned target degrees from a power law distribution. An edge is then placed between two nodes with a probability positively dependent on their distance and negatively dependent on their degrees. They investigated greedy routing with the distances on the ring as a means of navigating in the network.

Krioukov et. al [29] considered using a hyperbolic plane as the hidden social space. Nodes are uniformly distributed in a radius R disk in a hyperbolic plane with edges placed in pairs with distance smaller than r. They show that such a graph is naturally scale-free and that greedy routing with hyperbolic distance delivers the packets with high success rate.

2.2 Navigation in real-world networks Although the theoretical models and algorithms above are very inspiring, they require networks to satisfy certain properties, such as a scale-free degree distribution, or they require additional node-identifying information. They may fail in real-world networks that violate these properties, and are impossible to apply in real-world networks that lack node-identifying information, such as anonymous social networks and non-social networks. Even if such information is available, the edges in real-world networks do not necessarily follow the distribution of similarity-dependence specified in these theoretical models. These algorithms have been tested on only a few real-world networks for which node-identifying information was available, and without much success. For example, the hierarchical organization model [19] has been shown to work well only on the HP email network, for which messages were delivered in a median of four hops, but perform poorly for the Club Nexus online social network due to incomplete data or less structured hierarchy [1] (even using extensive profile knowledge the local search has a medium of 21 steps and a mean of 53 steps). Using geographical locations has shown only a delivery rate of 13% to deliver a message to the target city (not the target individual!) on a LiveJournal data set [32]. Kleinberg's small world model [27] is used to fit the 'web of trust' of the email cryptography tool Pretty Good Privacy (PGP) [38]. But the delivery rate is only 32% delivery rate with mean 26 steps.

Compared with prior work, we do not require the network to be scale-free as in [2, 24, 6]. We do not assume nodes stay in a given space as in [27, 6, 29]. And we do not require any node identities or geographical locations as in [27, 19, 25, 39]. Instead of requiring an embedding of the network in an observed space, our task is to discover the hidden space of real-world networks, which is possibly specific to each network. As it turns out, with this hidden space discovered, greedy routing achieves *much higher* success rates compared with previous experiments with real-world spatial embedding.

2.3 Graph embedding and greedy routing Embedding a graph in Euclidean spaces with small metric distortion has been studied actively in recent years. It is known that any graph of n vertices can be embedded in $\mathbb{R}^d$ with $d = O(\log n)$ such that the graph distance is distorted by a factor of at most $1 + \varepsilon$, for any $\varepsilon > 0$, with the Euclidean distance in the embedding (please refer to the book [33] for a large body of work on this topic). It can be shown that the Internet Autonomous Systems network (AS-network) can be embedded in $\mathbb{R}^7$ such that most of the routes can estimate their inter-delay fairly accurately by their Euclidean distances [36, 26]. Similar efforts have also been done for transportation networks [9].

The greedy routing algorithm minimizing 'distance' to the destination is used pervasively in ad hoc wireless network routing [8, 23], which is the motivation for the study of graph embeddings that supports greedy routing schemes. Most existing work only considered embedding in low dimensional Euclidean space such as $\mathbb{R}^2$ or $\mathbb{R}^3$ [31, 37].

3 Embedding and Greedy Routing

In this section we describe the algorithms we used for embedding a given complex network in Euclidean space and greedy routing with the 'social distances' in that space. We considered embedding in both Euclidean spaces and hyperbolic spaces. We remark that these algorithms are 'off the shelf' techniques. However, even these simple methods produce extremely short paths with greedy routing. We discuss the implication and empirical significance of the results in a later section.

3.1 Embedding in Euclidean Spaces Multi-dimensional scaling (MDS) is a classical method for embedding a set of nodes in $\mathbb{R}^d$. It takes an $n \times n$ distance matrix P as input, outputs an $n \times d$ coordinate matrix such that the ℓ_2 distance between any pair of nodes approximates the corresponding distance in P. MDS is done as follows:

1. Transfer $P = (p_{ij})_{n\times n}$ to $B = (b_{ij})_{n\times n}$ s.t.

$$b_{ij} = -\frac{1}{2}(p_{ij}^2 - \frac{1}{n}\sum_{j=1}^{n} p_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n} p_{ij}^2 + \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} p_{ij}^2).$$

   This step shifts the matrix to the center by subtracting the mean.

2. Perform eigen-decomposition on B s.t.

$$B = VAV^T,$$

   in which matrix A is a diagonal matrix with the eigenvalues ordered from largest to smallest and matrix V contains the corresponding eigenvectors. Set

$$X = VA^{1/2}.$$

   Then X gives the coordinate of the n nodes in n dimensional Euclidean space.

3. To reduce the embedded dimension to d, take the largest d eigenvalues from A and their corresponding eigenvectors in V to form an $n \times d$ coordinate matrix.

MDS has a running time of $O(n^3)$ and requires $O(n^2)$ space, once the distance matrix P is given. In our case, the distance matrix will include the all pairs shortest path distances.

For large data sets, MDS is too slow. V. de Silva and J. Tenenbaum proposed a landmark based multi-dimensional scaling method (LMDS) [16]. LMDS selects a small subset of nodes as landmarks, and performs MDS to embed these landmarks. The other nodes will measure their distances to the landmarks. By performing distance-based triangulation, each node will embed itself. Denoting by n the number of nodes, k the number of landmark nodes, d the embedded dimension, the procedure is as follows:

1. Let P_k denote the squared distance matrix of landmark nodes, compute B_k by adopting step1 in MDS, which shifts P_k to its center.

2. Let λ_i and $\vec{v_i}$ denote the ith largest eigenvalue and its corresponding eigenvector of B_k, respectively. Compute

$$w_i = \vec{v_i}^T/\sqrt{\lambda_i},$$

   then $W = (w_1, w_2, \ldots, w_k)$ is the transformation matrix of triangulation.

3. Let δ_i, $i = 1, \cdots, k$, be the hop distance vector from node i to the k landmarks, taking its mean as

$$\delta_\mu = \sum_{i=1}^{k} \delta_i/k.$$

4. Given a node i, the embedded coordinate

$$\vec{x_i} = -\frac{1}{2}W(\delta_i - \delta_\mu),$$

   which is an affine transformation of the distance vector. The first d elements of $\vec{x_i}$ gives the desired embedding result.

LMDS only requires $O(kn)$ space and its running time is $O(dkn + k^3)$, where k is the number of landmarks, d is the dimensionality to be embedded to. Since $d < k \ll n$, LMDS requires much less space and time than MDS. In our applications k and d are both chosen to be small constants. Thus LMDS has a linear running time.

Another benefit of LMDS is that it does not require the knowledge of the entire network. Note that we only need to know the shortest path distances from the landmarks to all other nodes in the network. This can be achieved by using breadth-first search from the landmarks with a running time of $O(kn)$.

MDS requires the knowledge of the entire network and thus is less desirable for a distributed system, compared with LMDS. To use landmark MDS, each landmark can flood the network so every node knows the distance to the landmarks. One of the landmark nodes performs the classic MDS method on the pairwise distance matrix $M_{k\times k}$ on all landmarks and broadcasts the landmark coordinates to the entire network. Non-landmark nodes then perform distance-based triangulation on their own to derive their coordinates. Since we typically use a constant number of landmarks. k can be considered as a constant, so MDS on the landmark nodes takes a constant amount of time and space. Alternatively, we may also use the distributed MDS [15] to compute the coordinates of the landmark nodes in the first step.

3.2 Embedding in Hyperbolic Spaces A hyperbolic plane is a 2D Riemannian manifold with negative curvature. A popular model is the Poincaré disk model, in which points are in a unit disk, and the straight lines are segments of circles contained in the disk orthogonal to the boundary of the disk, or else diameters of the disk. If u, v are two vectors with Euclidean norm less than 1, we define an isometric invariant by

$$\delta(u,v) = 2\frac{||u-v||^2}{(1-||u||^2)(1-||v||^2)},$$

where $||\cdot||$ denotes the usual Euclidean norm. Then the distance function is $d(u,v) = \operatorname{arccosh}(1+\delta(u,v))$.

R. Kleinberg [28] proposed a method to embed a graph in hyperbolic plane. First we compute a spanning tree T of the graph. The tree T is then embedded in hyperbolic plane so that each vertex is given a coordinate. Note that although one can route on the tree edges only, the non-tree edges are also used by greedy routing to produce (hopefully) shorter paths.

If the maximum degree of the tree T is d, there will be at most d branches from the root. We take a regular polygon P such that each side corresponds to a branch. Take any side of P, we introduce a hyperbolic isometry τ which maps endpoints of this side to 1 and -1 while mapping the midpoint of its corresponding arc on the boundary circle to $-i$. Define $u = \tau(0)$ as the point of root on the Poincaré disk, then the virtual coordinate of root node r can be computed as $\mu_r^{-1}(u)$. μ_r is a Möbius transformation defined below. Consider two hyperbolic isometries

$$a : z \longmapsto -z,$$

$$b : z \longmapsto \tau(\rho(\tau^{-1}(z))), \rho(z) = e^{2\pi i/d}z,$$

the Möbius transformation is computed from a and b. For each pair of parent and child nodes $(p(w), w)$, if the virtual coordinates of $p(w)$ is known as $f(p(w))$, points on the Poincaré disk for $p(w)$ and w are u and v, respectively, then $f(w) = \mu_w^{-1}(v)$. This process can be made to work in a distributed manner and it takes $O(n)$ time to find coordinates for all nodes. For the scheme described above the length of the coordinates can be exponential in n in the worst case. There are techniques to reduce the length of the coordinates to be $O(\operatorname{polylog} n)$.

3.3 Greedy Routing in Latent Spaces After the construction of the coordinate system, greedy routing will be used to navigate from source to destination, by sending the message to a neighbor closer to the destination. The distance is evaluated by the Euclidean distance or hyperbolic distance, depending on the embedding. When a node does not have a neighbor closer to the destination than itself, greedy routing gets stuck and fails to deliver the message. The performance is evaluated by measuring the success rate of greedy routing and the average routing path length.

Greedy Routing with Degree Information. A small world network typically has a power law degree distribution [4], with many low-degree nodes and a few high degree 'hubs'. As high degree nodes have more neighbors, routing to high degree nodes has a higher chance of encountering a node with the destination as immediate neighbor. This is the idea used in the degree-based greedy routing by Adamic et al. [2]. However, one drawback of pure degree-based routing is that messages arriving at hub nodes have no direction to 'come down' to low degree destinations.

We propose a hybrid scheme by using greedy routing based on both distance and degree. Apart from the reason that high degree nodes are 'more connected' to other nodes, they also tend to be embedded nearer to the 'core' of the network than low degree nodes, as will be shown later with our experiments. We set a degree threshold T. The nodes with degree higher than T are the *hubs* of the network. The rest of the nodes are at the *periphery* of the network. When the message is at a periphery node s, we send the message to its neighboring node with highest degree. If s does not have a neighbor with higher degree or it is already a hub, we go back to greedy routing with the embedded distances.

4 Experimental Results

In this paper we tested our method on a number of complex networks, including five real-world networks we were able to find from publicly available sources, except some really small size networks. We also tested on three artificially graphs generated from models. The details of these data sets are shown in Table 1.

The astrophysics collaboration network (ASTRO) [35] is the network of coauthorship between scientists posting preprints on the Astrophysics E-Print Archive between January 1, 1995 and December 31, 1999.

The Internet AS network (AS) [14] is a snapshot of autonomous systems generated by The Cooperative Association for Internet Data Analysis (CAIDA). This snapshot is taken on March 11, 2009. We do not specify the type of links, say, as customer-provider links or peer-peer links. So AS is an undirected graph with only connectivity information.

The Gnutella network (Gnu) is a data set used by Adamic in [2]. Gnutella was a popular P2P application, the data set is a small world network with hubs as the high degree nodes.

The actor network (ACT) [3] is a data set extracted from imdb, each line between two nodes means those two actors collaborated in one movie.

The NYT news network (NYT) puts an edge between individuals that appear together in at least two articles(strong

Data set name	Num(node)	Num(edge)	Avg degree	Num(landmark)
Astrophysics Coauthor Network (ASTRO)	14,845	119,652	16.1202	100
Internet AS Network (AS)	31,277	70,527	4.5098	100
Gnutella Network (Gnu)	574	835	2.9178	30
Actor Network (Actor)	171,427	6,984,461	81.4861	100, 300 and 500
NYT Network (NYT)	209,158	666,956	6.378	100 and 500

Table 1: Empirical Data Sets

juxtapositions) in the New York Times newspaper from 1981 to 2007.

The preferential attachment model (Pre) [3] assumes a dynamic network with nodes coming in one by one. When a new node joins the network, it is connected to existing nodes with probability proportional to their current degree. The preferential attachment model introduces the 'rich gets richer' hypothesis in the formation of complex networks and has been a popular model in explaining natural systems. In this experiment, we take $14,845$ nodes, each additional node has outdegree 8.

The Boguna-Krioukov-Claffy model (BKC) [6] is a recently proposed model assuming a hidden metric space on which the nodes reside. In particular, the hidden space is assumed to be a ring with the nodes uniformly placed on the ring. Each pair u and v has a distance $|uv|$ on the ring. Each node u is assigned a target degree k_u, drawn from a power law distribution. Each edge (u, v) is added with probability proportional to

$$(1 + |uv|/(k_u k_v))^{-\alpha},$$

where α is a model parameter. Intuitively the probability that two nodes have an edge is inversely dependent on their distance in the hidden space, and positively dependent on their target degrees. In general nodes near to each other in the hidden space are more likely to be connected. If they are hub nodes, they are also likely to be connected even though they are far away in the hidden space. In this experiment, we draw the degree from power law distribution Prob$\{k = i\} \sim 1/i^3$, which is the degree distribution in the preferential attachment model. α is set to be 2.

The Erdos-Renyi model (ER) [21] is the uniform random graph model, in which each pair of nodes is connected uniformly randomly. It has a small diameter. We include this graph as a baseline, as a random graph has no special structure to help with navigation.

For all the theoretical model networks, we vary the average degree to examine the performance dependency on the number of edges.

4.1 Embedding by MDS and LMDS We evaluate the performance of embedding with MDS and landmark MDS. LMDS is computationally more efficient than MDS, yet it gives similar embedding results even when k is small. Figure 1 shows the comparison between MDS and LMDS. Since the coordinates of landmark nodes by LMDS are consistent with those in MDS [16], LMDS with all nodes as landmarks is exactly MDS. The testing graph is generated by means of the preferential attachment model, Boguna-Krioukov-Claffy model and Erdos-Renyi models with $2,000$ nodes and $8,000$ edges. The embedded dimension is 6. Given a pair of nodes u and v, denote the distance generated by MDS and LMDS as $d_{u,v}$ and $d'_{u,v}$, respectively. The *individual distortion* of LMDS relative to MDS is $\rho'_{u,v} = |d_{u,v} - d'_{u,v}|/d_{u,v}$. The *average distortion* is $\rho' = \sum_{u,v} \rho'_{u,v}/n^2$. The average distortion drops as the number of landmarks decreases. Figure 1 also shows that by selecting only a small number of landmarks, high success rates for greedy routing can nevertheless be achieved.

4.2 Greedy routing results on Euclidean space We evaluated two strategies for routing: greedy routing, and greedy routing with degree information. We also evaluated two different strategies for selecting landmarks, random selection and selection of high degree nodes. The left four plots of Figure 2 shows the success rate for different embedded dimensionality. The Internet autonomous system network has the highest success rate, which reaches nearly 80% with an embedded dimensionality of 50. The preferential attachment graph achieves up to 56% success with hubs selected as landmarks, while the astrophysics collaboration network has a success rate of more than 60% with random landmark selection. The average path length of the AS network and preferential attachment graph stays around 4, and ASTRO network stays around 5. This result shows that greedy routes are short. The results also show that a random graph does not have good navigability. The graph generated by the recently proposed BKC model, though claimed to be navigable in [6], does not work better than the preferential attachment model or the real networks. Among the model networks, random graph are the most 'unstructured' one. It does not have power-law degree distribution or high clustering coefficient. Our experiments show that random graphs are indeed less navigable than all the other networks.

Figure 2 also shows that for some networks selecting highest degree nodes as landmarks performs better, while

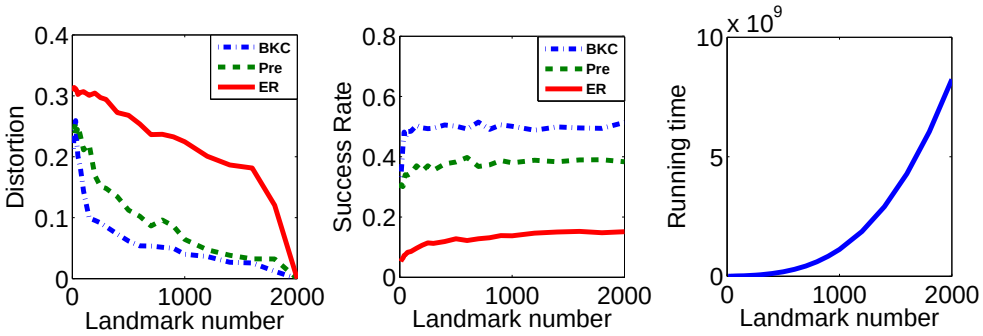

Figure 1: Performance comparison of LMDS and MDS. BKC,Pre and ER stands for Boguna-Krioukov-Claffy model, Preferential Attachment model and Erdos-Renyi model, respectively.

for other networks random selection works better. Detailed analysis shows that ASTRO network contains multiple clusters, while most of the highest degree nodes belong to only one cluster. Therefore, selecting highest degree nodes cannot capture the true network structure, which leads to lower success rate than random selection. Figure 3 (i) also shows that random selection works better in the actor network.

In the right two plots of Figure 2, we set the high degree threshold such that 20% of the nodes are hubs respectively. Contrary to the intuition that greedy routing with degree information improves the performance, we do not see much improvement. This shows that degree information is not important in our method, but rather that the distances in the discovered space truly help with greedy routing.

Success rate is possibly influenced by landmark number and embedded dimensionality. While Figure 2 shows that the success rate grows with dimension, we take a larger network to see how those two factors influence the success rate.

The actor network is a large network with more than $170,000$ nodes. Figure 3 (i) shows an apparent growth of success rate as the embedded dimensionality grows. Landmark number does not seem to play an important role, since using 100, 300, and 500 landmarks gives similar curves. Remark that the number of dimensions must be smaller than the number of landmarks. Similar results are obtained for NYT network, which contains similar number of nodes with actor network, but with much fewer edges. It suggests that network density does not play an essential role in our greedy routing. Despite that actor network and NYT network are much larger than our other data sets, the average path lengths are below or around 5 for all experiments.

Since the empirical networks have different average degrees, it is essential to justify whether average degree is an important factor in our embedding and greedy routing. Figure 3 (ii) shows the impact of average degree, it takes three model networks preferential attachment model, BKC model and Erdos-Renyi Model. Combining with Figure 2, we show how success rate and average path length changes while the average degree changes from 16 to 8 and 4. The result turns out that average degree does not show an essential impact on our embedding and routing.

4.3 Greedy routing results on hyperbolic space As shown in [28], greedy routing on hyperbolic space is proved to have guaranteed success. Therefore the length of average path length is the crucial performance factor to evaluate. In hyperbolic embedding, different choices of spanning tree and the root of the spanning tree will lead to different routing pathes between source and destination.

We first use the shortest path spanning tree (SPT) rooted as the node with highest degree as the tree in the hyperbolic embedding. This tree is computed by flooding from the root. In the first phase, an arbitrary node will flood all the other nodes to compute the highest degree node in the network. Ties are broken arbitrarily. In the second phase, The node with maximum degree will perform flooding to get a breadth-first tree.

In the second method, we use random walk to generate a spanning tree (RWT). In particular, we take an arbitrary node u as the starting node, and uniformly randomly select one of its neighbors v. If v has not been visited, edge (u, v) is a tree edge. We move to v and perform the same strategy until all nodes are visited. The running time of this distributed process depends on the cover time of the random walk on the particular graph.and ranges from $O(n \log n)$ to $O(n^3)$. RWT is a spanning tree uniformly randomly selected from the set of all possible spanning trees.

Table 2 shows the average distortion, average path length and shortest path length for hyperbolic routing on the empirical networks. It shows that there is a huge difference between different spanning trees. The spanning tree selected by shortest path tree rooted on highest degree node achieves much smaller path distortion. We conjecture that the reason might be that SPT usually divides nodes into more balanced branches than the trees obtained from random walk.

5 Discussion

Prior work has demonstrated that under certain special conditions, networks can be navigated in a decentralized fashion with algorithmic methods. How navigable real-world

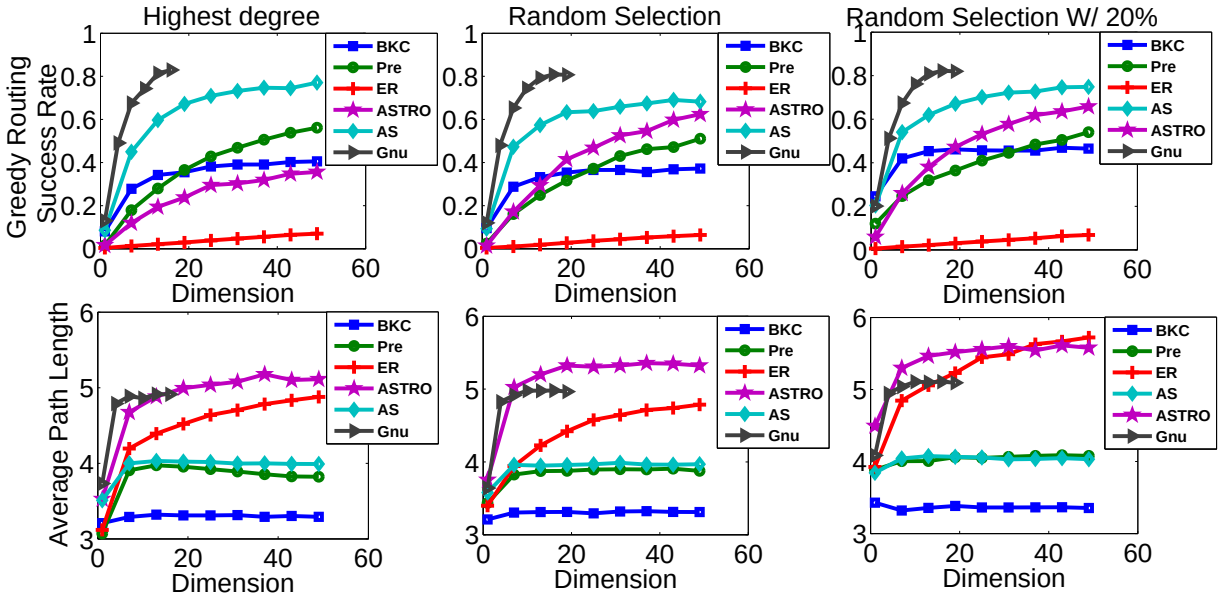

Figure 2: Simple greedy routing. Random Selection W/ 20%: landmark nodes are random selected, 20% highest degree nodes are hub nodes. Node number and average degree of ER, Pre and BKC are the same as ASTRO network. For abbreviations, please refer to Table 1 and Figure 1.

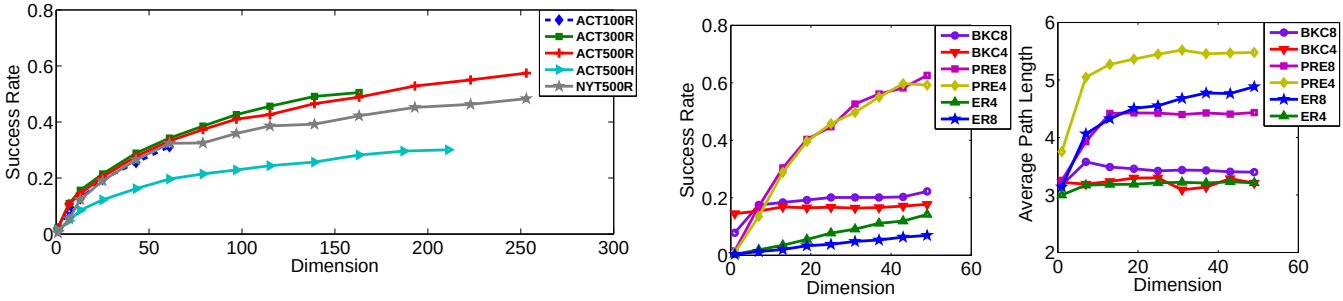

Figure 3: (i) Impact of landmark number and dimension. ACT500H is ACTOR network with 500 highest degree nodes as landmarks. (ii) Impact of average degree. BKC8 means a BKC network with average degree to be 8.

Data set name	SPT			RWT		
	Avg SPL	Avg RL	Avg distortion	Avg SPL	Avg RL	Avg distortion
ASTRO	4.579	5.837	1.459	4.579	23.080	5.770
AS	3.847	4.265	1.422	3.847	9.712	3.237
Gnutella	4.812	5.533	1.383	4.812	9.202	2.301
NYT	5.270	5.971	1.994	5.271	21.846	4.368
Actor	4.185	5.617	1.404	4.182	27.990	6.997
Pre	3.318	4.001	1.334	3.318	18.012	6.004
BKC	3.252	3.927	1.309	3.252	20.425	6.808
ER	4.863	7.978	1.995	4.863	36.270	9.067

Table 2: Greedy routing with embedding in hyperbolic space. For abbreviations, please refer to Table 1 and Figure 1. SPL and RL stands for shortest path length and routing path length, respectively.

networks that may violate these special conditions are, has remained largely unknown. We have shown here that five actual empirical networks, social and non-social, of very diverse origin and topology, can all be embedded in some hidden space such that effective navigation becomes possible. Using rather elementary methods, we deliver a majority of packages in under six hops in every network. Even more surprisingly, navigation results are generally better, not worse, than in simulated model networks.

Our experimental results yielded a number of interesting findings.

Greedy routing performance. The performance of navigation in real-world networks using a hidden space improves substantially on previous results. Past small-world experiments in which human participants forwarded messages in social networks had much lower delivery rates due to uncooperative participants. Milgram's experiment used node identifying information about the destination such as name, gender, occupation, etc. 64 out of 296 letters arrived at the destination (around 20% success rate), with an average path length of around 5.5. Dodds et. al [17] conducted the same Milgram's experiment with emails instead of snail mails and extended the geographical range to the whole world. They show that only 1.5% messages arrived at the destination, as many participants dropped from the game. Liben-Nowell et. al [32] conducted a greedy routing simulation on a LiveJournal data set, thus eliminating user participation issues. They used geographical information as the greedy routing criterion (a message is delivered to a friend closer to the target geographically) and considered delivery successful if the message arrived at the city where the destination individual resides (not the actual precise location of the individual target). Their simulations show a 13% success rate. We compare our results with Liben-Nowell's results, as also their investigation involves real data sets and users are assumed cooperative. Liben-Nowell's results use only geographical location as the greedy routing criterion, which may only partially capture the attributes attributed to navigability. We use the coordinates extracted from the embedding. Our results confirm that using a proper embedding into latent spaces greatly helps with navigation in the network. It remains as interesting future work to see whether the embedded space corresponds to a set of real-world attributes or combination of attributes.

Adamic [2] proposed degree routing and performed greedy routing simulations on the same Gnutella data set. The degree routing strategy selects the highest degree node among the neighbors which have not been explored, if all neighbors have been visited, routing process enters a dead end and fails. Results on comparison between degree-based routing strategy and our method is shown in table 3. Although degree routing gives very good success rate on dense model networks, average path length is too high to be considered as practical. Besides that, as average degree of Pre, BKC and ER reduces below 4, success rate drops quickly to below 40%, while average path length is still higher than 100.

Sandberg [38] tried to fit a particular small world model by J. Kleinberg [27] to a real data set, the web of trust, so as to discover the users' locations on a grid using Markov Chain Monte Carlo method. The method has only demonstrated limited success with 32% delivery rate and mean 26 steps. Part of the reason could be that it is not clear whether J. Kleinberg's model truly reflects the structure of the real data set.

There is a tradeoff between the dimensionality of the embedding and the success rate of greedy routing using that embedding. On one hand, the success rates grows as the dimensionality is increased, but the growing speed slows down. Thus increasing the embedding dimensionality has diminishing return. Depending on the network structure and the size of network, some networks, for example AS and BKC, have a tipping point on the growth of success rate. On the other hand, the size of the coordinates for each node grows linearly as the embedding dimensionality. The embedded dimension should also be strictly smaller than the number of landmarks, hence one may have to use more landmarks with a higher dimensionality.

Real networks are more navigable than models. In this paper we did not try to pinpoint the exact properties that make certain networks navigable, like many small-world models that have been proposed were intended to capture. Rather, we show that even with off the shelf techniques one is able to get very reasonable performance on *real-world* networks, something we never expected before we ran these experiments. An interesting additional discovery is that the performance of real-world networks turns out to be better, not worse, than equivalent model networks with the same network size and average degree. That is, real networks are more navigable than existing models! Besides results shown in the paper, we also tested the J. Kleinberg's model [27]. With around $14,000$ nodes and average degree 8, it gives low success rate(around 20%) and average path length is around 20. This is not good compared to greedy routing using Kleinberg's grid coordinates, which guarantees delivery and gives similar average path length. In a sense, this suggests that there is some structural feature to real-world complex networks that has not yet been captured by any single theoretical model. It is also possible that hybrid models that combine the different navigation-facilitating characteristics of various small-world models would do better. How to build such hybrid models is an interesting line of future work.

The choice of latent space and embedding. We experimented with both Euclidean spaces and hyperbolic spaces as

	ASTRO	AS	Actor	NYT	Gnutella	Pre	BKC	ER
Success Rate	29.7%	35.7%	74.9%	26.8%	51.8%	99.7%	98.4%	98.9%
Avg Path Length	55.30	10.96	572.54	341.55	14.02	206.97	338.00	631.58

Table 3: Degree-based routing

the hidden spaces. Embedding in both spaces achieves high success rate and more importantly, very low greedy routing path length. This raises the question whether the topology of the hidden space matters after all, and whether there is some structure of real-world networks that transcends the hidden spaces.

What is the optimal hidden space? – results on the BKC model. The BKC model [6] assumes a hidden space for the nodes in a social network but does not mention how to *discover* this space. The model assumes a ring structure as the hidden space and greedy routing is based on the distance on the ring. With this method, routing on the BKC graph with $14,845$ nodes and $118,325$ edges has a success rate of about 25% with average path length to be 8.07. As Figure 2 shows, our method will give around a 47% success rate with the average path length to be 3.36. Thus, paradoxically, although the BKC model constructed the network with an assumed hidden space, the assumed hidden space is apparently not the optimal one. It remains as interesting future work how to characterize the 'hidden space' from the network structure and how to find the optimal hidden space, if there is one.

6 Conclusion and future work

Our experiment is the first to systematically investigate the conjecture made in earlier small world navigation studies that many real-world complex networks are navigable. That is, it is possible to discover a hidden metric space purely from the network connectivity information alone that permits greedy routing on the coordinates in the hidden space to discover extremely short paths for a majority of node pairs. We confirm the conjecture, delivering packages in a majority of cases in each of our empirical networks. The five networks we consider were not handpicked, representing radically different contexts and showing diverse topologies. In addition to confirming the navigability hypothesis, they show the success of relatively straight-forward embedding and navigation methods. Likely, the true navigability of these networks is even better.

Acknowledgement. We acknowledge Lada Adamic for sharing with us the Gnutella network data set generated by Clip2. We thank Steve Skiena and Charles Ward for sharing with us the New York Times data. X. Ban and J. Gao would acknowledge partial support from NSF through CNS-0643687.

References

[1] L. Adamic and E. Adar. How to search a social network. *Social Networks*, 27:187–203, 2005.

[2] L. A. Adamic, R. M. Lukose, A. R. Puniyani, and B. A. Huberman. Search in power-law networks. *Physics Review E.*, 64:046135, 2001.

[3] A. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[4] M. Barthelemy and L. Amaral. Small-world networks: Evidence for a crossover picture. 82(15), 1999.

[5] E. S. Bogardus. Social distance in the city. *Proceedings and Publications of the American Sociological Society*, 20:40–46, 1926.

[6] M. Boguna, D. Krioukov, and K. C. Claffy. Navigability of complex networks. *Nature Physics*, 5:74–80, January 2009.

[7] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications*. Springer-Verlag, 2nd edition, 2005.

[8] P. Bose, P. Morin, I. Stojmenovic, and J. Urrutia. Routing with guaranteed delivery in ad hoc wireless networks. *Wireless Networks*, 7(6):609–616, 2001.

[9] U. Brandes, F. Schulz, D. Wagner, and T. Willhalm. Generating node coordinates for shortest-path computations in transportation networks. *Journal of Experimental Algorithmics*, 9(1), 2004.

[10] R. S. Burt. *Structural Holes: The Social Structure of Competition*. Cambridge University Press, 1992.

[11] V. Buskens. *Social Networks and Trust*. Kluwer, 2002.

[12] V. Buskens and A. van de Rijt. Dynamics of networks if everyone strives for structural holes. *American Journal of Sociology*, 114:371–407, 2008.

[13] C. T. Butts. Predictability of large-scale spatially embedded networks. In *Dynamic Social Network Modeling and Analysis*, pages 313–323, 2003.

[14] CAIDA. The caida as relationships data,090311. http://www.caida.org/data/active/as-relationships/, 2009.

[15] J. A. Costa, N. Patwari, and A. O. H. III. Distributed weighted-multidimensional scaling for node localization in sensor networks. *ACM Transactions on Sensor Networks*, 2(1):39–64, 2006.

[16] V. de Silva and J. Tenenbaum. Sparse multidimensional scaling using landmark points. Technical report, Stanford University, 2004.

[17] P. Dodds, M. Roby, and D. Watts. An experimental study of search in global social networks. *Science*, 301:827, 2003.

[18] G. Dooms, Y. Deville, and P. Dupont. Constrained metabolic network analysis: discovering pathways using cp(graph), 2005.

[19] D.Watts, P.Dodds, and M.Newman. Identity and search in social networks. *Science*, (296):1302–1305, 2002.

[20] D.Watts and S.Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684):409–410, 1998.
[21] P. Erdos and A. Renyi. On random graphs. *Publicationes Mathematicae*, 6:290–297, 1959.
[22] H. Jeong, S. Mason, A.-L. Barabasi, and Z. Oltvai. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.
[23] B. Karp and H. Kung. GPSR: Greedy perimeter stateless routing for wireless networks. In *Proc. of the ACM/IEEE International Conf. on Mobile Computing and Networking (MobiCom)*, pages 243–254, 2000.
[24] B. Kim, C. Yoon, S. Han, and H. Jeong. Path finding strategies in scale-free networks. *Physical Review E*, 65(027103), 2002.
[25] J. Kleinberg. Small-world phenomena and the dynamics of information. In *In Advances in Neural Information Processing Systems 14*. MIT Press, 2001.
[26] J. Kleinberg, A. Slivkins, and T. Wexler. Triangulation and embedding using small sets of beacons. In *Proc. 45th IEEE Symposium on Foundations of Computer Science*, pages 444–453, 2004.
[27] J. M. Kleinberg. The small-world phenomenon: an algorithm perspective. In *Proc. of ACM Symposium on Theory of Computing (STOC)*, pages 163–170, 2000.
[28] R. Kleinberg. Greedy routing using hyperbolic space. pages 1902–1909, 2007.
[29] D. Krioukov, F. Papadopoulos, M. Boguna, and A. Vahdat. Greedy forwarding in scale-free networks embedded in hyperbolic metric spaces. In *ACM SIGMETRICS Workshop on Mathematical Performance Modeling and Analysis (MAMA)*, June 2009.
[30] B. Latane, J. H. Liu, A. Nowak, M. Bonevento, and L. Zheng. Distance matters: Physical space and social impact. *Personality and Social Psychology Bulletin*, 21:295–805, 1995.
[31] T. Leighton and A. Moitra. Some results on greedy embeddings in metric spaces. In *Proc. of the 49th Annual Symposium on Foundations of Computer Science*, pages 337–346, October 2008.
[32] D. Liben-Nowell, J. Novak, R. Kumar, P. Raghavan, and A. Tomkins. Geographic routing in social networks. In *Proceedings of the National Academy of Science*, volume 102, pages 11623–11628, 2005.
[33] J. Matoušek. *Lectures on Discrete Geometry*. Springer, 2002.
[34] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.
[35] M. Newman. Astrophysics collaborations. In *Proc. Natl. Acad. Sci.*, volume 98, pages 404–409, 2001.
[36] E. Ng and H. Zhang. Predicting Internet network distance with coordinates-based approaches. In *Proc. IEEE INFOCOM*, pages 170–179, 2002.
[37] C. H. Papadimitriou and D. Ratajczak. On a conjecture related to geometric routing. *Theor. Comput. Sci.*, 344(1):3–14, 2005.
[38] O. Sandberg. Distributed routing in small-world networks. In *Proc. of Algorithm Engineering and Experiments (ALENEX)*, 2006.
[39] O. Simsek and D. Jensen. Navigating networks by using homophily and degree. In *Proceedings of the National Academy of Sciences*, pages 12758–12762, September 2008.
[40] J. Travers and S. Milgram. An experimental study of the small world problem. *Sociometry*, 32:425, 1969.
[41] A. van de Rijt, X. Ban, and R. Sarkar. Effective networking when connections are invisible: Comment on reagans and zuckerman. *Industrial and Corporate Change*, 17:945–952, 2008.
[42] R. Williams, E. Berlow, J. Dunne, A. Barabasi, and N. Martinez. Two degrees of separation in complex food webs. *Proceedings of the National Academy of Sciences*, 99(20):12913–12916, 2002.