CS224W Project Writeup: On Navigability in Small-World Networks

Danny Goodman

12/11/2011

1 Introduction

Graphs are called *navigable* one can find short paths through them using only local knowledge [2]. This requires a very precise link distribution, and yet many important real-world networks are navigable [1]. How does such a link distribution arise?

Oskar Sandberg proposed a model of network formation in an attempt to explain how navigability arises [2]. The model takes a geographic nearest-neighbor network and adds links based on a model of intrinsic similarity between nodes. The model produces the right link distribution in special theoretical cases and simulations, but it's full theoretical consequences are unknown. Sandberg conjectures that the model in general behaves as it does in his special case. The main difficulty in the proof is that links are not created independently, so it is not easy to prove statements about an entire greedy path.

In this project, I attempted to solve this difficulty for a special case. I did not succeed, but present theoretical and computational evidence in favor of this conjecture.

2 The Interest Model and Double Cluster Graphs

This section defines the precise theoretical setting of the project, which is a special case of Sandberg's model [2].

Definition A *Double Cluster Graph* is a set of vertices $\{x_i\} = V$ along with two independent distance functions, d_1 and d_2, which we think of as 'geographic' and 'intrinsic' distance. An edge (x_i, x_j) is in the graph if, for all $x_k \in B_{d_1(x_i,x_j)}$, $d_2(x_i, x_k) \geq d_2(x_i, x_j)$.

Intuitively, a node is connected to every node that is the most intrinsically similar node within it's geographic neighborhood. This means that close geography and intrinsic similarity both increase the likelihood of links. Intrinsic distance can represent similar interests in a social network, similar content in

the web graph, or any other trait that leads geographically far nodes to form links.

We are interested in families graphs where the interest metric d_2 is randomly chosen, called Random Double Clustering Graphs. It will be helpful to control the randomness through a random permutation π, so that $d_2(x, y) = d(\pi(x), \pi(y))$, where d is the same interest metric over every instantiation of the random graph in a given family.

Sandberg's conjecture is as follows:

Conjecture 1. (Sandberg) *Greedy paths in randomized double cluster graphs with bounded doubling dimension have expected size of* $O(\log n)$.

This conjecture is important because navigability is an extremely important, and yet non-robust, feature of all the major real-world networks mentioned in class. It is not well understood why diverse networks exhibit this property. The best way to understand a phenomenon is to construct the simplest, most general possible model which corresponds to reality and also exhibits the phenomenon. This conjecture takes a large step in that direction. Sandberg [2] proves this conjecture for a simplified model of a family of Directed Double Cluster Graphs, and provides some computational evidence for the main conjecture.

2.1 Sandberg's Special Case

Sandberg is only able to show navigability for a special family of graphs, called *Directed Double Cycle Graphs*, defined by the following metrics:

$$\begin{aligned} d_1(u, v) &= v - u \mod n \\ d_2(u, v) &= \pi(v) - \pi(u) \mod n \end{aligned}$$

where π is a uniformly chosen permutation on n elements. The above distance functions are not symmetric, so the resulting double cluster graph is a directed graph – and a cycle graph as well due to the modulus. While this configuration is highly artificial, Sandberg simulates a few more realistic double cluster graphs without this restriction and finds the same scaling of path lengths, up to a constant factor. He then conjectures that the same scaling holds for all double cluster graphs.

If true, this conjecture would provide the only simple analytical static (non-evolving) model to explain why the precise inverse rank relation holds in real networks. The importance of this has already been discussed. Sandberg's formulation imposes very little restriction on the heterogeneity of the double graph – only a property called 'bounded doubling' which relates the volume of a node-set to it's diameter. It is therefore not obvious that this conjecture holds in its full generality. It does seem clear from Sandberg's simulations, however, that the conjecture holds true over a far greater space of double cluster graphs than Sandberg is able to demonstrate in his paper.

2.2 Undirected Double Cycle DiGraphs

The main difficulty in proving Sandberg's conjecture in the general case is that the link-formation process is not independent. One must therefore find path-independent bounds on greedy step size (which our simulations support), or find ways to argue about entire paths. We study the simplest special case that faces the independence problem: the *Undirected Double Cycle DiGraph* on n nodes, defined by the following metrics:

$$\begin{aligned} d_1(u,v) &= \min(u-n-v, u-v, u+n-v) \\ d_2(u,v) &= d_1(\pi(u), \pi(v)) \end{aligned}$$

For shorthand, we write $d_1(u,v) = |u-v|$. These distance metrics are clearly symmetric, so each cycle is undirected, but the graph is a digraph because the interest-based link-forming process does not necessarily produce reciprocal edge pairs.

Unlike the Directed Double Cycle, greedy paths in the Undirected Double Cycle do not need to monotonically approach their target in either metric. They may overshoot the target and backtrack with respect to d_1, and are not even required to approach the target in d_2.

3 Cracking the Conjecture

The general proof is expected to proceed along these lines:

How then to sharpen this bound? I am attempting a 2-pronged approach:

1. Study the independence problem more deeply. My intuition says an $O(1)$ halving bound should be possible. Try to be clever.
2. Failing that, study simulations of the undirected cycle search problem. Is $O(1)$ really the right bound? Which theoretical bounds from the directed case hold in the simulated undirected case? Can we prove those?

The independent cycle problem is a logical first step towards the whole conjecture. It must introduce new ideas which may be useful in the general case, and it is likely an easier problem. The same joint theoretical and computational approach can then be applied to the full conjecture.

4 Theoretical Evidence for the Conjecture on Undirected Double Cycle Digraphs

We imagine a full proof of Sandberg's conjecture would proceed along these lines:

1. Show that a greedy step of the search process on expectation reduces the physical distance to target by a fixed fraction. This is usually easy given the $\log n$ degree and link distance profile.

2. Prove that the above result is independent of any previous steps in the search path. This is the difficult step in the general case, and necessary to place a global bound of $O(\log n)$ steps.

We prove the first step in this section, but the second step is elusive.

Lemma 1. *Consider a node x in a large undirected double cycle graph G on n nodes. Let x_i be a node such that $|x_i - x| = i$. Denote by A_i the event that there is an edge $x \to x_i$. Then A_i and A_j are independent for $i \neq j$.*

Consider a node $x_j \in (x, z)$ such that $|x - x_j| = j$. Note that there are $2j-2$ nodes strictly closer to x than x_j, so the chance of an edge from x to x_j is $\frac{1}{2j-1}$. This is because any of these nodes is equally likely to be the most intrinsically similar.

Assume that $i < j$. Fix a permutation $\pi_0 : X_0 \to [1, 2i-1]$ of the elements $X_0 = \{x_i\} + \{y \mid 0 < |y - x| < i\}$ such that $\pi_0(x_i) = 2i - 1$. Each occurrence of A_i defines a permutation π_0. Now let π_1 be a permutation on $X_1 = \{y \mid 0 < |y - x| < j\}$ that preserves the order of $\pi_0(X_0)$ (ie $\pi_0(u) < \pi_0(v) \to \pi_1(u) < \pi_1(v)$ etc). A_i also defines a single π_1.

Since $|X_1| = 2j-2$, there are 2j-1 unique ways to insert x_j into the permutation, all of which are equally likely. Therefore, $P(A_j|A_i) = \frac{1}{2j-1} = P(A_j)$. $\square$

Theorem 1. *Consider a very large undirected double cycle graph G on n nodes. Consider the greedy path from node x to node z, $x, x_1, x_2, \ldots, x_m, z$ and let $d(x, z) \geq 2$. The first step in this path reduces expected distance to z at least $\frac{k}{20}$.*

Despite that the constant is not optimal, the significance of this bound is that it leads to logarithmic greedy paths if it holds in general. We explain this after the proof.

Assume w.l.o.g that $0 < x < z < n/2$. In fact, we further assume that $z < \frac{n}{4}$, so that we only have to consider greedy paths to one side of x. Note that the expected first step lengthens if we consider paths to both sides of x, but the analysis is more difficult.

Denote by B_j the event that there is no edge to j distance or closer of z, for $j \leq k - 2$:

$$\begin{aligned} Pr(B_j) &= \prod_{i=k-j}^{k+j} (1 - Pr(A_j)) \\ &= \prod_{i=k-j}^{k+j} \frac{2i-2}{2i-1} \end{aligned}$$

Note that when $j = k-1$ there is such an edge with probability 1 by x's nearest neighbor. We will need a bound on this quantity:

$$\begin{aligned} Pr(B_j)^2 &= \prod_{i=k-j}^{k+j} \left(\frac{2i-2}{2i-1}\right)^2 \\ &> \prod_{i=k-j}^{k+j} \frac{2i-3}{2i-1} \\ &= \frac{2k-2j-3}{2k+2j-1} \end{aligned}$$

What is that chance that there is an edge from x to exactly distance j from z? The edge could fall on either side of z:

$$\frac{1}{2k-2j-1} + \frac{1}{2k+2j-1} - \frac{1}{(2k-2j-1)(2k+2j-1)} = \frac{4k-3}{(2k-2j-1)(2k+2j-1)}$$

Let x_1 be the first node on the greedy path from x to z. Then for $j \in (1, k-1)$ we have

$$Pr(|x_1 - z| = j) = \frac{Pr(B_{j-1})(4k-3)}{(2k-2j-1)(2k+2j-1)}$$

These facts allow us to compute the desired lower bound. In the following calculations our aim is only to achieve a linear estimate, so we make estimates that sacrifice the constant coefficient:

$$E\Big[|x-z| - |x_1 - z|\Big] = \frac{k}{2k-1} + \sum_{j=1}^{k-1}(k-j)Pr(|x_1 - z| = j)$$

The rather ugly left-hand expression is necessary, instead of the more intuitive $|x - x_1|$, because x_1 may be on either side of z from the perspective of x, so long as it is closer to z.

$$\begin{aligned} (k-j)Pr(|x_1 - z| = j) &> \frac{(k-j)(4k-3)}{(2k-2j-1)(2k+2j-1)}\sqrt{\frac{2k-2j-1}{2k+2j-3}} \\ &\geq \frac{\cancel{(2k-2j-1)}(2k-1)}{\cancel{(2k-2j-1)}(2k+2j-1)}\sqrt{\frac{2k-2j-1}{2k+2j-3}} \\ &> \frac{\left(k-j-\frac{1}{2}\right)^{3/2}}{(2k)^{3/2}} \\ \sum_{j=1}^{k-1}(k-j)Pr(|x_1 - z| = j) &> \frac{1}{(2k)^{3/2}}\int_0^{k-1} j^{3/2} dj \\ &= \frac{1}{(2)^{3/2}}\frac{2(k-1)}{5}\left(1-\frac{1}{k}\right)^{3/2} \end{aligned}$$

We also assume $k \geq 2$ because $x_1 = z$ with probability 1 if $k = 1$. Finally, then, we obtain:

$$E\Big[|x - z| - |x_1 - z|\Big] > \frac{k}{20}$$

□

If this relation could be shown to hold for all steps in a greedy path, it would give the following bound on path size s:

$$E(s(x, z)) < 1 + \frac{\log |x - z|}{\log 20 - \log 19}$$

However, the above proof makes use of specific on-average facts which do not necessarily hold in the context of previous steps in a greedy path.

5 Numerical Evidence for the Conjecture on Undirected Double Cycle Digraphs

This section describes computational studies suggesting that the above $O(k)$ bound on the greedy hop holds for successive hops, independent of preceding hops. The code submitted as supplementary materials creates random Undirected Double Cycle Digraphs with 1,000 nodes each, randomly samples start and end nodes, and saves the greedy path between each node pair.

With 10,000 paths as samples, the following least squares fit confirms the overall log-size paths:

$$s(x, z) \approx 0.962 + 1.186 * |x - z|$$

See figure 1 for a plot. The fit intercept is close to the expected value of 1.

The purpose of the simulation is to test the independence problem: whether subsequent hops on a greedy path are somehow systematically shorter, which could jeopardize the desired routing time. We performed least squares regressions of step size *step* on remaining distance to target (*rem*), and on *rem* and number of previous steps (*num*):

$$\begin{aligned} step &\approx -2.227 + 0.642 * rem \\ step &\approx -3.361 + 0.649 * rem + 0.256 * num \end{aligned}$$

These results show that number of previous steps has little effect on step size. The similarity of the *rem* coefficient in both regressions also supports this. Interestingly, the negative intercept in both regressions suggests that the early greedy steps are comparatively a bit longer than the later steps.

While this strongly suggests the independence of step expectation and previous path, it unfortunately does not suggest a theoretical plan of attack.

6 Conclusion

We have presented theoretical and numerical evidence that $O(\log n)$ greedy routing holds in Sandbergs model despite the independence problem. A full proof was elusive due to lack of imagination, but the conclusion appears highly likely. With some more efforts, this beautifully simple model may be shown to explain how navigability arises across so many classes of real networks.

References

[1] Jon Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symosium on Theory of Computing*, pages 163-170, 2000.

[2] Oskar Sandberg. Double clustering and graph navigability. In *Proceedings of CoRR*, 2007.

[3] Olof Mogren, Oskar Sandberg, Vilhelm Verendel, and Devdatt Dubhashi. Adaptive dynamics of realistic small-world networks. `http://arxiv.org/pdf/0804.1115v1`

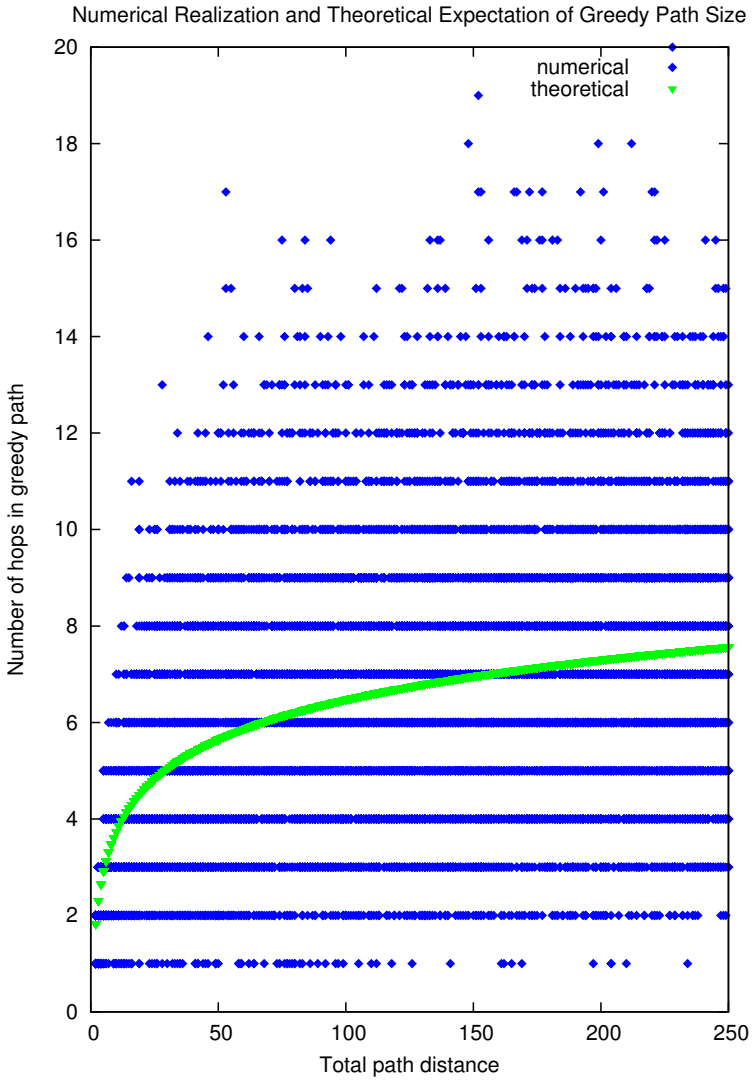

Figure 1: Path size as a function of path length.