Approximating Shortest Paths in Spatial Social Networks

Carlo Ratti
ratti@mit.edu

Christian Sommer
csom@mit.edu

***Abstract*—We evaluate an algorithm that efficiently computes short paths in social networks by exploiting their spatial component. The main idea is very simple and builds upon Milgram's seminal social experiment, where target individuals were found by having participants forward, or *route*, messages *towards* the target. Motivated by the somewhat surprising success of this experiment, Kleinberg introduced a model for spatial social networks, wherein a procedure called *greedy routing* can be used to find short, but not necessarily *shortest* paths between any two individuals.**

We extend Kleinberg's greedy routing procedure to explore $k \geqslant 1$ links at each routing step. Experimental evaluations on social networks obtained from real-world mobile and landline phone communication data demonstrate that such adaptations can efficiently compute accurate estimates for shortest-path distances.

I. Introduction

In the famous *small-world experiment* of Milgram [1], people in Nebraska were asked to send a letter to a person in Massachusetts. Participants were supposed to forward the letter to a person they knew on a first-name basis (defining a *social network*). Surprisingly, the letters that arrived at the destination were forwarded along chains of only six persons. Due to Milgram's experiment, we know the "six degrees of separation" phenomenon as the phenomenon that (almost) any two individuals are connected by a path of length six.

Even more remarkable is the phenomenon that such paths can actually be found by individuals using local information only. Kleinberg [2], [3], in his seminal work on small worlds, noted that social networks are *navigable*, meaning that short paths can be found using local information. Kleinberg suggested and analyzed an algorithm, called *greedy routing*, to find short paths in social networks with a spatial component (meaning, in our case, that each node has Euclidean coordinates): the algorithm at each node follows a link in the social network such that the Euclidean distance to the target is minimized.

II. Generalized Greedy Routing

We evaluate a simple shortest-path algorithm that leverages the spatial component of social networks. This algorithm efficiently computes (approximate) shortest paths between pairs of nodes *without* any preprocessing. The algorithm is parametrized as follows: instead of following only *one* link that minimizes the Euclidean distance (as in greedy routing), the algorithm follows the *top–k* links with respect to Euclidean distance.

The inherent tradeoff of this algorithm is between *accuracy* and *query time*: the more time the algorithm spends (larger k), the better the quality of the resulting path.

Similar techniques (known as A*) have previously been used successfully for road networks [4] with an inherent spatial structure. Our experimental results suggest that the geographical component can potentially be quite useful for computations in more complex networks such as social networks as well.

III. Experimental Evaluation

We experiment using large real-world network datasets with a spatial component, derived from Call Detail Records (CDR).

a) Mobile Portugal: Based on anonymized call data records stemming from the Orange mobile network in Portugal, a network on 1,822,756 nodes (corresponding to users) and 11,367,729 edges (corresponding to reciprocal phone connections) was extracted [5]. Most users were assigned to one out of 6,509 different cell towers as their "home base," computed as the cell tower that they happened to be using most frequently around midnight.

b) Landline UK: Based on an anonymized set of landline telephone calls in Great Britain, a network with a largest connected component of 20,704,523 nodes (corresponding to phone numbers, which could be associated with an individual, a household, or even an office) and 85,719,483 edges (corresponding to reciprocal phone connections) was extracted [6]. A node's geographic location is specified at the level of sub-regional switching facility groups.

Each network was loaded into main memory, together with its corresponding latitude and longitude coordinates (unfortunately, we have well-defined locations only for a subset of the nodes and some of these locations may be wrong). Then, we performed greedy routing / shortest-path queries for thousands of random source–target pairs.

Results and Interpretation

1) Greedy Routing ($k = 1$): In these experiments, the search was truncated as soon as the target *location* was reached (even though there are multiple nodes per location). We list the distance distributions in terms of graph distance (number of hops) if the target location was found (greedy routing may sometimes terminate in a dead end, where there is no outgoing link that improves upon the Euclidean distance).

For *Mobile Portugal*, the search was successful in 336 out of 1,000 (s, t) pairs, where we define a successful search

when the target location is found for GREEDY(s,t) **or** for GREEDY(t,s). We list the smaller distance in Table I. (Exhaustive search (BFS), was successful for 963 pairs.)

distance	0	1	2	3	4	5	6	7	8	9
# pairs	2	10	58	85	72	58	30	13	4	4

TABLE I: Graph distance distribution of bidirectional greedy $\min\{\text{GREEDY}(s,t), \text{GREEDY}(t,s)\}$ for the *Mobile Portugal* network.

For *Landline UK*, the success rate of the greedy routing is comparable (roughly 31%). Results are listed in Table II.

distance	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
# pairs	14	44	152	188	185	109	78	32	19	6	5	1			1	1

TABLE II: Graph distance distribution of bidirectional greedy $\min\{\text{GREEDY}(s,t), \text{GREEDY}(t,s)\}$ for the *Landline UK* network.

2) Shortest-Path Queries: The algorithm corresponds to greedy routing with $k \geqslant 1$ for the first phase. As soon as the target location (cell) has been reached, the algorithm switches to BFS for the second phase. We provide the mean values and standard deviations (among at least 5,000 queries) of the speedup relative to the unidirectional version of BFS, and of the multiplicative *stretch* relative to a shortest path. This multiplicative stretch is defined as the length of the resulting path divided by the shortest-path length. The running times are measured in terms of the number of settled nodes and the speedup is computed accordingly.

k	success	stretch	speedup
1	34.27%	1.643 ± 0.362	95.08 ± 1,162.07
2	74.14%	1.549 ± 0.439	78.35 ± 841.86
3	86.19%	1.478 ± 0.331	55.40 ± 611.93
4	91.17%	1.446 ± 0.278	34.97 ± 351.53
5	91.78%	1.426 ± 0.253	38.63 ± 356.43
6	91.36%	1.399 ± 0.238	29.74 ± 231.22
7	91.40%	1.387 ± 0.231	17.43 ± 138.82
8	91.22%	1.378 ± 0.233	9.78 ± 48.25
9	90.94%	1.372 ± 0.246	10.08 ± 50.53

TABLE III: Success rate, stretch, and speedup for k–GREEDY(s,t) on *Mobile Portugal* and various sizes of the envelope (parameterized by an integer k, specifying how many outgoing links to consider at each step).

As expected, the success rate of the algorithm increases with the envelope width k and the stretch decreases with the envelope width. For larger k, the closer we get to A*/BFS, the better the stretch. The average stretch is quite low and the worst-case stretches observed were also very reasonable, being 3, 3.67, 3.33, 2.25, 2.33, 2.2, 2.25, 2.2, and 2.25 for $k = 1, 2, \ldots 9$, respectively.

The speedups range from several orders of magnitude to even slower than BFS, resulting in a high variance (see last column of Tables III and IV). We observe that the target

k	success	stretch	speedup
1	31.56%	1.475 ± 0.317	106.51 ± 1,436.13
2	68.84%	1.408 ± 0.396	151.32 ± 2,861.16
3	76.38%	1.368 ± 0.296	82.47 ± 1,351.98
4	76.64%	1.330 ± 0.275	74.31 ± 1,326.08
5	77.53%	1.320 ± 0.254	114.96 ± 2,222.98
6	77.96%	1.296 ± 0.248	65.73 ± 957.11
7	78.44%	1.290 ± 0.246	58.24 ± 562.37
8	78.00%	1.290 ± 0.239	74.22 ± 1,663.19
9	77.63%	1.277 ± 0.247	73.30 ± 1,065.93

TABLE IV: Success rate, stretch, and speedup for k–GREEDY(s,t) on *Landline UK* and various sizes of the envelope (parameterized by an integer k, specifying how many outgoing links to consider at each step).

location is found quickly (Table V) but reaching the actual target (switch to BFS) requires our current implementation to visit many additional nodes. Some locations are popular in the sense that they "host" many nodes (some locations have more than 6K nodes). It might be possible to address these shortcomings by determining the locations of a node in a more fine-grained way.

Target locations are found quickly: Table V lists the distance distributions of greedy routing, following the k best outgoing links at each step. For example, for $k = 4$, the target *cell* is found after $h = 6$ hops for all but nine pairs, after exploring at most $k^h = 4\,096$ nodes (out of 1.8M nodes). Note that the running time is not necessarily bounded by $O(k^h)$, since there may be nodes with high degrees (say d), for which a straightforward implementation would spend time $O(\min\{kd, d \log d\})$ per node. For high-degree nodes, one might organize their neighbors in a data structure supporting efficient nearest neighbor queries.

k	0	1	2	3	4	5	6	7	8	9	10	11
1	2	10	58	85	72	58	30	13	4	4		
2		5	79	173	186	153	68	42	15	6	3	2
3		7	97	267	249	139	66	21	9	1		
4		7	90	259	220	106	31	7	1	1		
5		8	95	278	223	87	28	5				
6		6	78	237	158	62	13	2				

TABLE V: Distance distribution (# pairs at distance) of bidirectional greedy $\min\{k\text{–GREEDY}(s,t), k\text{–GREEDY}(t,s)\}$ for the *Mobile Portugal* network, *not* switching to BFS.

REFERENCES

[1] S. Milgram, "The small world problem," *Psychology Today*, vol. 1, pp. 61–67, 1967.

[2] J. M. Kleinberg, "Navigation in a small world," *Nature*, vol. 406, p. 845, 2000.

[3] ——, "The small-world phenomenon: an algorithm perspective," in *32nd ACM Symposium on Theory of Computing (STOC)*, 2000, pp. 163–170.

[4] R. Sedgewick and J. S. Vitter, "Shortest paths in Euclidean graphs," *Algorithmica*, vol. 1, no. 1, pp. 31–48, 1986, announced at FOCS 1984.

[5] F. Calabrese, Z. Smoreda, V. D. Blondel, and C. Ratti, "Interplay between telecommunications and face-to-face interactions: A study using mobile phone data," *PLoS ONE*, vol. 6, 2011.

[6] C. Ratti, S. Sobolevsky, F. Calabrese, C. Andris, J. Reades, M. Martino, R. Claxton, and S. H. Strogatz, "Redrawing the map of Great Britain from a network of human interactions," *PLoS ONE*, vol. 5, 2010.