COMS E6998-2: Analysis of Social Information Networks — Jan 31, 2011

Lecture 1 – Small World

Instructor: *Augustin Chaintreau* — Scribes: *Aditi Bhatnagar,Ari Golub*

1 Algorithmic Small world analysis

Milgram's experiment has shown that it is possible in a network of acquaintance for a person to find a shortest path to a target, even if this person knows only the general information about the target. Greedy routing" mechanism is introduced according to which, whenever a person v needs to route a message to a target t,he or she chooses among their acquaintances (i.e. the nodes connected to them by an edge), the node that is the closest to the target for the distance defined by the lattice.

1.1 Random uniform augmentation

Consider a randomly augmented lattice with dimension $k = 1$ and $q = p = 1$.

Proposition 1. *In a randomly augmented lattice of dimension $k = 1$ with N nodes, greedy routing uses in expectation at least $\frac{1}{4}\sqrt{N}$ steps.*

Proof. Let us consider a target t, and the neighborhood subset

$$I_l = \{\ u \in V\ \mid |u - t| \leq l\ \}\ .$$

This subset contains t and at most $\lfloor 2l \rfloor$ nodes. Starting from an initial point s, greedy routing construct a path $s = U_1, U_2, U_3$ by following at each step either a local edge in the grid or a shortcut. $(U_i)_{i\geq 1}$ are then random points in the grid. Since shortcuts are chosen independently of each other we can apply the principle of deferred decision, which states that previous random choices used by the algorithm do not impact the outcome of random choices not currently used. Let X_i be the node that the shortcut rooted at U_i points to. X_i forms an i.i.d. sequence of uniformly chosen points. Each of them lies in I_l with probability $\frac{1}{N}\lfloor 2l \rfloor \leq \frac{2l}{N}$. The probability that one of the n first elements of X_i lies in I_l is then upper bounded by the union bound:

$$\mathbb{P}\left[\bigcup_{i=1,\ldots,n} \{\ X_i \in I_l\ \}\right] \leq \sum_{i=1,\ldots,n} \mathbb{P}\left[X_i \in I_l\right] \leq \frac{nl}{N}\ .$$

Hence, for $n = l = \frac{1}{2}\sqrt{N}$, this occurs with probability at most $1/2$. We conclude that, with probability at least $1/2$ the n first shortcuts encountered $\{\ X_1, X_2, \ldots, X_n\ \}$ lie outside I_l. In that case, if we assume that s itself is not in I_l, the greedy procedure needs either

- to make n steps

- otherwise, to reach before t before n steps, it requires to reach t from the boundary of I_l using l local edges.

In both cases, $\frac{1}{2}\sqrt{N}$ steps are required.

Observation 2. *Hence the expected number of steps needed by greedy routing is lower bounded by a constant time the square root of N.*

□

Proposition 3. *In a randomly augmented lattice of dimension $k \geq 1$ containing N nodes, greedy routing uses in expectation at least $\frac{1}{4}N^{\frac{1}{k+1}}$ steps.*

Observation 4. *It may seem that the shortcut augmentation of the lattice does connect points better as k increases. But in a lattice of dimension k with N nodes, the distance in the lattice is of the order of $O(N^{\frac{1}{k}})$, so that the* relative *improvement obtained with shortcuts augmentation actually becomes worse as k increases.*

1.2 Random augmentation with bias

Reasons of the failure of Milgram's experiment under the uniform augmentation are:

- About $\sqrt{N}$ shortcuts exist on average that will lead to the interval I_l when $l = \sqrt{N}$ but these shortcuts are uniformly distributed among all N nodes and, hence, it takes a lot of steps to find one of these starting from an arbitrary point.
- Since they are uniformly distributed, no algorithm can hope to make some progress by following other shortcuts, since moving anywhere does not improve his chance to find one.

Definition 5. *A norm is a function that assigns a strictly positive length or size to all vectors in a vector space, other than the zero vector. Likewise $\|u - v\|$ is the sum of the absolute values of the coordinates of the vector (u-v). It gives the number of steps to go from position u to position v using the edges of the lattice. Different types of norms are Euclidean norm, Taxicab norm, P-norm, Maximum norm, Zero norm.*

- *Euclidean norm is denoted by* $\|X\| = \sqrt{{X_1}^2 + {X_2}^2 \cdots + {X_n}^2}$
- *P-norm is denoted by* $\|X\|_p = (\sum_{i=1}^{N} |X_i|^p)^{\frac{1}{p}}$.
- *Maximum norm or Infinity norm is denoted by* $\|X\|_\infty = max(|X_1|, |X_2|, \ldots, |X_n|)$

Augmenting lattice with a bias: Shortcuts connecting people are biased. Even if you happen to have a friend outside of your usual circles of acquaintanceship, it is extremely more likely that this connection is much closer to you than a typical arbitrary person. On the other hand, it is not impossible that you have a connection with someone very distant, it is simply much less likely. Kleinberg proposed to modify the shortcuts so as to reflect this bias. A *random biased augmented lattice* of dimension k containing N nodes with bias parameter r is defined as follows:

- We assume $V = \{ (i_1, \ldots, i_k) \in \{1, 2, \ldots, L\}^k \}$, (note that $N = L^k$).

- Nodes are connected to all other nodes whose distance in the lattice is at most p (i.e. $v = (i_1, \ldots, i_k)$ and $v' = (i'_1, \ldots, i'_k)$ are connected if $|i_1 - i'_1| + \ldots + |i_k - i'_k| \leq p$).
- In addition, each node is connected to q others nodes chosen independently such that

$$\mathbb{P}\left[u \rightsquigarrow v\right] = \frac{\frac{1}{\|u-v\|^r}}{\sum_{w \neq u} \frac{1}{\|u-w\|^r}} \ .$$

In the probability describing the chance to connect u and v, the denominator only plays the role of a normalizing constant.

The parameter r, usually called the *clustering coefficient*, describes the bias used in the augmentation of the grid. A node that is twice far away from u than v may still be chosen but only with a probability 2^r times less.

- As r approaches 0 the ratio between these probability approaches 1. When $r = 0$ the distance plays no role of any sort and hence the grid is a uniformly augmented lattice.
- As r grows, the distribution tends to favor only immediate neighbors. As r goes to infinity we can then effectively assume that the grid is not augmented.

Impact on greedy routing performance By introducing bias in the augmentation of lattice we can conclude that not all positions are equal when it comes to find the target. Hence, it may be possible to use shortcuts not as a way to get to the immediate neighborhood of target t but as a way to move towards that direction. Hence increases the probability to find a good shortcuts.

Theorem 6. *When $r = k$, greedy routing uses in expectation at most $O(\ln(N)^2)$ of steps.*

- *When $0 \leq r < k$, for any p and q, then as n grows any decentralized algorithm uses in expectation at least $\Omega(N^{\frac{d-r}{d+1}})$*
- *When $r > k$, for any p and q, then as n grows any decentralized algorithm uses in expectation at least $\Omega(N^{\frac{r-2}{r-1}})$*

Proof. We will consider this proof for dimension k = 1. The proof of each case contains two parts: First, a bound on the normalizing constant used in the probability distribution of the shortcuts. Second, a study of the progress of greedy routing which uses this bound.

If $k = 1$:

$$\sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u-v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r} \ . \tag{1}$$

Wherever u is positioned in the line, it has at least one side (either left or right) which contains at least $N/2$ neighbors. For each value of $j = 1, \ldots, \lfloor N/2 \rfloor - 1$, it has one neighbor at distance j on this side of the line, which proves the lower bound. The upper bound is obtained after observing that u has at most 2 neighbors at distance j for all j and that the maximum distance cannot be more than N.

The case $r < 1$

$$\sum_{v \neq u} \frac{1}{\|u-v\|^r} \geq \sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \geq \int_1^{\lfloor N/2 \rfloor} \frac{1}{x^r} dx \geq \frac{1}{1-r} \left((\lfloor N/2 \rfloor)^{1-r} - 1 \right)$$

The second inequality comes from the fact that, as $x \mapsto \frac{1}{x^r}$ is a decreasing function, it is smaller than $\frac{1}{j^r}$ on the interval $[j, j+1]$. As a consequence, the sum used in the normalizing constant asymptotically grows polynomially, with coefficient $1 - r > 0$. In particular, for $N \geq 2^{\frac{3-r}{1-r}}$ we have that $(N/2)^{1-r} \geq 2$ and hence $(N/2)^{1-r} - 1 \geq \frac{1}{2}(N/2)^{1-r}$. We then deduce:

$$\text{For } \; N \geq 2^{\frac{3-r}{1-r}} \, , \sum_{v \neq u} \frac{1}{\|u-v\|^r} \geq c_1 N^{1-r} \; \text{ where } \; c_1 = \frac{1}{2(1-r)2^{(1-r)}} \, .$$

This proves that, however u and v are located, the probability that the shortcut originating in u leads to v is becoming small *polynomially* with N:

$$\mathbb{P}\left[u \rightsquigarrow v\right] \leq \frac{1}{c_1 N^{1-r}}$$

If we denote I_l as the set of nodes at distance at most l from the target,

$$I_l = \{ \; u \in V \; \mid |u - t| \leq l \; \} \; ,$$

then since the number of nodes in this subset is less than $2l$, the probability that a shortcut originated in u leads to a node in I_l is upper bounded by $\frac{2l}{c_1 N^{1-r}}$. We consider the sequence of nodes visited by the greedy routing procedure $U_1, U_2, \ldots, U_k$, and for each of them denote by X_i the destination of the shortcut originating at U_i. The probability that one of the n first elements of X_i lies in I_l is then upper bounded by the union bound:

$$\mathbb{P}\left[\bigcup_{i=1,\ldots,n} \{ \; X_i \in I_l \; \} \right] \leq \sum_{i=1,\ldots,n} \mathbb{P}\left[X_i \in I_l\right] \leq \frac{n2l}{c_1 N^{1-r}} \, .$$

Choosing $l = n = \lambda N^{\frac{(1-r)}{2}}$ the probability above is upper bounded by a constant independent of N. By choosing λ sufficiently small we have that it is less than $1/4$. This indicates that with probability at least $3/4$ all the n first shortcuts found by greedy routing connect with a node outside of I_l. Starting from a point s outside of I_l, greedy routing cannot succeed in finding the target t in less than $\min(n, l) = \lambda N^{\frac{1-r}{2}}$.

Observation 7. *Finding the target requires here either to use more than n steps or to traverse from the border of I_l to the target using only local edges. In expectation, greedy routing needs a number of steps at least* $(1/2)(3/4)\lambda N^{\frac{1-r}{2}}$.

Breaking the lower bound Is it possible to break the lower bound set by Milgram's experiment? To suggest that it is possible is a priori not true– it is only true if you are already near the target or land near the target at random. Otherwise, it could still take any number of steps. All points are not equal, so progress is possible but not necessary. Using different values of r gives you very different results as to

whether the lattice exhibits "small world" properties. The critical case, where $r = k$, a neighborhood of t of radius $\frac{d}{2}$ contains $\frac{d}{2}^k$ nodes, each of which may be chosen with probability roughly $\frac{1}{\frac{3d}{2}^k}$. The growth of the size of the ball (i.e. the radius) compensates for the decrease in probability. This is a harmonic distribution.

The case $r > 1$: When $r > 1$, the connections that are made are less random and appear closer to the origin node. As a result, our lattice starts to approach the design of a regular lattice. We begin to break the idea of "small world" and short paths do not exist– the algorithm needs $N^{\frac{r-k}{r-(k-1)}}$ steps.

As in the previous case, we wish to establish a negative result hence our goal will be to provide an upper bound on the chance to make sufficient progress. However, the argument will be different this time, as the main obstacle is that the probability of having sufficiently long shortcuts is not large enough to allow the greedy procedure to move towards the destination sufficiently fast.

Indeed, we know that any node u in the line has at most 2 neighbors with distance j in the lattice, and the series which characterizes the normalizing constant, as shown in Eq.(**??**) converges. The fact that the series converge indicates that we deduce now a bound on the probability of reaching all nodes that are sufficiently far.

$$\sum_{v \neq u, \|u-v\|>m} \frac{1}{\|u-v\|^r} \leq 2 \sum_{j=m+1}^{N} \frac{1}{j^r} \leq 2 \left(\int_m^N \frac{1}{x^r} dx \right) \leq \frac{2}{(r-1)m^{r-1}} .$$

The last inequality is obtained after replacing the integral on $[m, N]$ with the integral on $[m, +\infty]$, which can only make this bound looser, and computing its value.

Since the normalizing constant is always greater than 1 the inequality above implies that for any m the probability for any node u to be connected through a shortcut to a node at distance larger than m is less than $\frac{2}{(r-1)m^{r-1}}$. We now consider the n first shortcuts encountered by greedy routing, as made in the previous proof. Following the union bound, we can deduce that the probability that at least one of them connect two nodes at distance larger than m is less than n times the above probability (*i.e.* $\frac{2n}{(r-1)m^{1-r}}$).

Let us now assume that we can choose m and n in such a way that this probability is smaller than $1/4$, this would implies that with probability at least $3/4$ all the n first encountered shortcuts connect two nodes at distance at most m (a probability event we denote by $\mathcal{E}$). We may assume that the initial distance between s and t is at least $N/4$. This event occurs with a probability $1/2$ and hence in intersects the event $\mathcal{E}$ at least for a probability $1/4$.

When both event occur then in order to complete the walk from s to t, greedy routing requires at least $\min(n, N/4m)$ steps, since the first n steps of the walk has a length at most m. This would imply, in expectation, that number of steps needed by greedy routing is at least $\frac{1}{4}\min(n, N/4m)$.

Now in order to complete the proof, we need to show that we can choose n and m so that $\frac{2n}{(r-1)m^{r-1}} \leq 1/4$ and $1/4\min(n, N/4m)$ is large as N grows.

The first condition is satisfied as long as $n \leq \frac{1}{8(r-1)}m^{r-1}$. We can choose n to be exactly this value as making n large is only helping to satisfy the second ondition. Hence, both conditions reduces to finding m such that $\min(\frac{1}{8(r-1)}m^{r-1}, N/4m)$ is large as N grows.

The value of m has opposite role in order to maximize each term, so that intuitively this minimum will be the largest when the two terms have the same order. In particular, if we choose $m = N^{\frac{1}{r}}$ this minimum is a constant multiplied by $N^{\frac{r-1}{r}}$, which proves the result of the theorem.

The case r=1: Finally we are left with the only positive result, which occurs at the critical case. It is interesting to see first why the proof of the case $r < 1$ and $r > 1$ do not apply. First, for $r = 1$ the series characterizing the normalizing constant is the harmonic series, hence it does not converge and we cannot apply the previous argument bounding the probability to find large links. Also, as opposed to the case $r < 1$ the series does not grow as fast as a polynom, which explains why we cannot use this argument to show that all probability, independently of the position of u and v becomes small.

We first obtain a upper bound on the series, as we observed that it diverges not as fast as polynom. Indeed, when $r = 1$, following Eq.(**??**),

$$\sum_{v \neq u} \frac{1}{\|u - v\|} \leq 2(1 + \sum_{j=2}^{N} \frac{1}{\|j\|}) \leq 2(1 + \int_{1}^{N} \frac{1}{x} dx) \leq 2(1 + \ln(N)) \leq 2(\ln(3N)) \ .$$

This implies that for any u the probability that it it connected with a shortcut to node v is at least $1/(2\ln(3N)d(u,v))$.

Greedy routing, initially started in a point s constructs a chains of nodes visited $U_1, U_2, \ldots$ until it reaches t. Let us say that U_i is in phase j if we have $2^j \leq \|U_i - t\| \leq 2^{j+1}$. Since the initial distance is at most N, we now that U_1, the starting point of the walks is in phase j_0 with $j_0 \leq \ln(N)/\ln(2)$. Note also that, as greedy routing decreases the distance to the target at each step, the phase of this walk can only decreases with the number of steps made.

The core of the argument for the theorem is to show that each phase of this walk is short (*i.e.* it involves a logarithmic number of step). This will imply the result because there are also a small number of phase (*i.e.* a logarithmic number).

We first consider the following quantity: Given that U_i is in phase j, what is the probability that U_{i+1} is in phase $j' < j$? According to the definition, all nodes in phase $j' < j$ are those who are at distance at most 2^j from the target t. This contains at least 2^j nodes (the target t may be on the border of the line, but it has at least 2^j neighbors within this distance in one direction).

The key observation is that, for every node v in phase $j' < j$, since U_i is in phase j, the distance between U_i and v can be bounded by triangular inequality:

$$\|U_i, v\| \leq \|U_i, t\| + \|t, v\| \leq 2^{j+1} + 2^j \leq (3/2)2^{j+1} \ .$$

Hence, the probability that U_i has a shorcut leading to a node in phase $j' < j$ is at least

$$\sum_{v | \|v-t\| < 2^j} \frac{1}{2\ln(3N)(3/2)2^{j+1}} \geq \frac{2^j}{(2\ln(3N)(3/2)2^{j+1})} \geq \frac{1}{6\ln(3N)} \ .$$

In other words, for any step taken by greedy routing in phase j the next step will be in a smaller phase with probability at least $(6\ln(3))^{-1}$. Note that this event only depends on the shortcuts chosen at this step and hence, the shortcuts that will be visited in the next steps are independent from this event.

This implies that, if we denote by S_j the number of steps made by this walk inside phase j, we can bound the probability that $S_j \geq i$ geometrically, hence we have:

$$\mathbb{E}[S_j] = \sum_{i \geq 1} \mathbb{P}[S_j \geq i] \leq \sum_{i \geq 1} \left(1 - \frac{1}{6\ln(3N)}\right)^{i-1} = 6\ln(3N).$$

Assuming that greedy routing starts in phase j_0, the total number of steps it needs to reach t is $S_{j_0} + S_{j_0-1} + \ldots + S_1$. By the linearity of expectation, and since $j_0 \leq \ln(N)/\ln(2)$, we have that it takes in expectation less than $6/(\ln(2))\ln(3N)\ln(N)$, which is less than $c\ln^2(N)$ for some constant c, proving the result.

Analysis and summary Greedy routing performs at $O\ln^2(N)$, but it is possible to find paths as short as $\ln(N)$ with extra information about the nodes. Practically, we can observe a harmonic distribution using closeness of rank instead of distance to measure the tightness of the lattice. Milgram's experiment proves that social networks are navigable. Individuals can take advantage of short paths with basic information (social or economic information, for example). This is at odds with uniform random graphs that have no predictability. The key ingredients that explain navigability are having an easy space to route (e.g. grids, trees etc.) and a subtle harmonic augmentation (e.g. ball with radius).