Lecture 3: Structure (3)

How the friendship we form connect us? Why we are within a few clicks on Facebook?

COMS 6998: Social Networks
Monday September 22th
Milgram’s “small world” experiment

- It’s a “combinatorial small world”
- It’s a “complex small world”
- It’s an “algorithmic small world”
  - Beyond uniform random augmentation
Autopsy of “Small-world” failure

* In a uniformly augmented lattice shortcuts do exist
  o About $\sqrt{N}$ shortcuts leads to $I_l$ when $l = \sqrt{N}$.

* But they are dispersed among $N$ nodes

* Moreover, previous steps do not lead to progress
  o So need about $N/\sqrt{N} = \sqrt{N}$ trials

* Is there another augmentation?
The 10 papers that will make you a social expert
10 sociological must-reads

People “love those who are like themselves”, “Similarity begets friendship”
- Nichomachean Ethics, Aristotle & Phaedrus, Plato

Do you think homophily produces or hinder small world?

Homophily in Online Dating: When Do You Like Someone Like Yourself?

Andrew T. Fiore and Judith S. Donath
MIT Media Laboratory
20 Ames St., Cambridge, Mass., USA
{fiore, judith}@media.mit.edu

ABSTRACT
Psychologists have found that actual and perceived similarity between potential romantic partners in demographics, attitudes, values, and attractiveness correlate positively with attraction and, later, relationship satisfaction. Online dating systems provide a new way for users to identify and communicate with potential partners, but the information they provide differs dramatically from what a person might glean from face-to-face interaction. An analysis of dyadic interactions of approximately 65,000 heterosexual users of an online dating system in the U.S. showed that, despite these differences, users of the system sought people like them much more often than chance would predict, just as in the offline world. The users’ preferences were most strongly same-sexing for attributes related to the life course, like marital history and whether one wants children, but they also demonstrated significant homophily in self-reported physical build, physical attractiveness, and smoking habits.

NATURE OF ONLINE PERSONALS DATA
We analyzed data from one online dating system in particular. Through an agreement brokered by the Media Laboratory with an online dating Web site (the “Site”), we obtained access to a snapshot of activity on the Site over an eight-month period, from June 2002 through February 2003. The data included users’ personal profile information, their self-reported preferences for a mate, and their communications via the site’s private message system with other users. Anonymous ID numbers distinguished unique users.

Table 1 indicates which profile characteristics users could specify about themselves and about the partners they would like to meet.

Data about private messages exchanged by the users included the sender, recipient, subject, text, date and time of delivery, and whether the recipient had read the message.

Author Keywords
Online personals, attraction, computer-mediated communication, online dating, relationships

ACM Classification Keywords
H5.3. Group and Organization Interfaces; Asynchronous
Augmenting lattice with a bias

* What if the augmentation exhibits a bias
  o Most of the people you know are near,
  o Occasionally, you know someone outside

* Does this break the lower bound proof?

* Does finding a neighborhood of \( t \) becomes easier?
How to model augmentation bias

Formal construction:
1. Connect nodes at distance $p$ in a regular lattice
2. Connect each node to $q$ other nodes, chosen with a biased probability
3. $p=q=1$ to simplify

The small-world phenomenon: An algorithmic perspective.
How to model augmentation bias

* Formal construction:
  1. Connect nodes at distance \( p \) in a regular lattice
  2. Connect each node to \( q \) other nodes, chosen with a biased probability

\[
\Pr [u \leadsto v] = \frac{1}{\|u-v\|^r} \sum_{v \neq u} \frac{1}{\|u-v\|^r}
\]

* \( r \) may be called the **clustering coefficient**
  * If a node is twice further, probability is \( 2^r \) times less

The small-world phenomenon: An algorithmic perspective.
Impact of clustering coefficient

Small values of $r$
Approaches uniform augmentation

Large values of $r$
Approaches original lattice
Can we break the lower bound?

(a) Yes, finding a neighborhood of $t$ becomes easier

A PRIORI NOT TRUE

- It is easier only if you are already near the target
- In general, it can take a larger number of steps
Can we break the lower bound?

(b) Yes, for another reason
- All positions are not equal, hence progress is possible
- As shortcut are used recursively, probability increases
- So we need to study the sequence of progress
Augmented lattice (dimension k)

Navigable small world!
dist. alg need $O(\log^2(N))$ steps

Combinatorial Small world
(Short paths exist)
dist. alg. need $N^{(k-r)/(k+1)}$ steps

Not a small world
(Short paths do not exist)
alg. need $N^{(r-k)/(r-(k-1))}$ steps

Theorem 1.
- When $r = 1$, greedy routing uses in expectation at most $O(\ln(N)^2)$ of steps.
- When $0 \leq r < 1$, for any $p$ and $q$, then as $n$ grows any decentralized algorithm uses in expectation at least $\Omega(N^{\frac{1-r}{2}})$
- When $r > 1$, for any $p$ and $q$, then as $n$ grows any decentralized algorithm uses in expectation at least $\Omega(N^{\frac{r-1}{r}})$
Can we break the lower bound?

* Example: $k=1$ (augmented lines) with $r=1$
  - If we start from $u$ such that distance $d(u,t)=l$
  - How many nodes are in interval $I_{l/2}$?
  - For $v$ in $I_{l/2}$, can we lower bound $P[u\rightarrow v]$?
Can we break the lower bound?

* Example: \( k=2 \) (augmented lattice)
  - If we start from \( u \) such that distance \( d(u,t)=l \)
  - How many nodes are in interval \( I_{l/2} \)?
  - For \( v \) in \( I_{l/2} \), can we lower bound \( P[u->v] \)?
The critical case

- Assume $r = k$ (dimension of the grid)
  - A neighborhood of $t$ of radius $d/2$
  - Contains $(d/2)^k$ nodes
  - Each may be chosen with probability roughly $1/(3d/2)^k$
  - Growth of ball compensates probability decreases!

- Harmonic distribution.

The small-world phenomenon: An algorithmic perspective.
Theorem 1.  

- When $r = 1$, greedy routing uses in expectation at most $O(\ln(N)^2)$ of steps.

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

- When $r > 1$, for any $p$ and $q$, then as $n$ grows any decentralized algorithm uses in expectation at least $\Omega(N^{\frac{r-1}{r}})$
We assume dimension $k=1$

For each case ($r<1$, $r>1$, $r=1$) we use two steps

- **STEP1**: Obtain a bound on the probability

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

- **STEP2**: Characterize the progress of greedy routing
  Introducing nodes visited, denoted $U_1, U_2, \ldots, U_i$
  the destination of shortcuts, denoted $X_1, X_2, \ldots, X_i$
We have when dimension $k=1$:

$$\sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r}$$
We have when dimension $k=1$:

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

- We can then understand why $r=1$ is a critical case
- $r<1$ : series diverge (roughly as $N^{1-r}$ for large $N$)
- $r>1$ : series converge
Case r<1 (1/3)

**STEP1:** \[ \sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r} \] (divergent series)

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

- When N is sufficiently large this implies

For \( N \geq 2 \frac{3-r}{1-r} \), \[ \sum_{v \neq u} \frac{1}{\|u - v\|^r} \geq c_1 N^{1-r} \] where \( c_1 = \frac{1}{2(1-r)2^{(1-r)}} \).
Case r<1 (2/3)

* **STEP1:** \[
\sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r}
\]

  o Which implies \[
\sum_{v \neq u} \frac{1}{\|u - v\|^r} \geq \sum_{j=1}^{\lfloor N/2 \rfloor - 1} \frac{1}{j^r} \geq \int_1^{\lfloor N/2 \rfloor} \frac{1}{x^r} dx \geq \frac{1}{1 - r} (\lfloor N/2 \rfloor)^{1-r} - 1)
\]

  o Hence, we have whenever N is large

\[
\mathbb{P} [u \sim v] \leq \frac{1}{c_1 N^{1-r}}
\]
* STEP2: Use the same proof as the uniform augmented lattice

- We introduce \( I_l = \{ u \in V \mid |u - t| \leq l \} \)

- We observe

\[
P \left[ \bigcup_{i=1,\ldots,n} \{ X_i \in I_l \} \right] \leq \sum_{i=1,\ldots,n} P[X_i \in I_l] \leq \frac{n2l}{c_1N^{1-r}}.
\]

- Choosing \( l = n = \lambda N^{(1-r)/2} \) this probability is <1/4

- Concluded that expected # steps is at least \( \text{cst} \ N^{(1-r)/2} \)
STEP 1:

\[
\sum_{j=1}^{[N/2]-1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r} \quad \text{(convergent series)}
\]

- Bound P[v is at least distance m from u] faraway

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

* **STEP2:** $P[u$ connected to $v$ at distance $> m]< \frac{2}{(r-1)m^{r-1}}$
  
  o $P[X_1$ or $X_2$ or ... or $X_n$ with distance $> m]< \frac{2n}{(r-1)m^{r-1}}$
  
  o $P[X_1$ and $X_2$ and ... and $X_n$ distance $< m]>$
    
    we call this event “small shortcuts”
  
  o On this event, if initial distance $> N/4$ then greedy routing needs at least min($n, N/4m$) steps
  
  o Choosing $m=N^{1/r}$ and $n=(2/4(r-1))N^{(r-1)/r}$ ensures that “small shortcuts” has proba $> \frac{3}{4}$ hence that the expected steps in greedy routing is at least $\text{cst}^* N^{(r-1)/r}$
**CASE r=1 (1/3)**

∗ **STEP1:**  
\[ \sum_{j=1}^{[N/2]-1} \frac{1}{j^r} \leq \sum_{v \neq u} \frac{1}{\|u - v\|^r} \leq 2 \sum_{j=1}^{N} \frac{1}{j^r} \]  
(slowly divergent)

○ In particular we can bound the normalizing constant

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

○ So that \( P[u \text{ connected to } v] \) at least

\[ \frac{1}{\|u - v\|} \times 2 \ln(3N) \]
**Case r=1 (2/3)**

* STEP1: \( P[u \text{ conn. } v] \geq \frac{1}{||u - v|| \cdot 2 \ln(3N)} \)

* STEP2: Let \( U_1, U_2, \ldots, U_j \) be nodes visited by the walk
  - We say \( U_i \) is “in phase \( j \)” if \( 2^j \leq ||U_i - t|| \leq 2^{j+1} \)
  - \( U_1 \), the starting point, is in phase \( j_0 \leq \ln(N)/\ln(2) \)
  - Let \( S_j \) the # of steps of greedy routing in phase \( j \)
  - Total number of steps: \( S_{j_0} + S_{j_0-1} + \ldots + S_1 \)

  - We only need to bound \( E[S_j] \) for all \( j \)
Case r=1 (2/3)

* STEP1: \( P[u \text{ conn. } v] \geq \frac{1}{||u - v|| \cdot 2 \ln(3N)} \)

* Main argument
  - Assuming \( U_i \) in phase \( j \), how likely is \( U_{i+1} \) in phase \( j' < j \)?
    - At least \( P[U_i \text{ connected to } v \text{ such that } |v-t| \leq 2^i] \)
  - There are at least \( 2^j \) nodes satisfying \( |v-t| \leq 2^i \)
  - Each of them is at distance at most \( (3/2)^{2^{j+1}} \) from \( U_i \)
  - Hence, we have \( \sum_{v: ||v-t|| < 2^j} \frac{1}{2 \ln(3N)(3/2)2^{j+1}} \geq \frac{2^j}{(2 \ln(3N)(3/2)2^{j+1})} \geq \frac{1}{6 \ln(3N)} \cdot \)
  - This implies \( E[S_j] \leq 6 \ln(3N) \) proving the results
The critical case

- Assume $r=k$ (dimension of the grid)
  - A neighborhood of $t$ of radius $d/2$
  - Contains $(d/2)^k$ nodes
  - Each may be chosen with probability roughly $1/(3d/2)^k$
  - Growth of ball compensates probability decreases!

- Harmonic distribution.

The small-world phenomenon: An algorithmic perspective.
Theoretical follow ups

* Is the analysis of greedy routing tight?
  o Yes, greedy routing performs in $\Omega(\log^2 n)$

* Can we find path as short as $\log(n)$ (shortest path)?
  o Yes, with extra information on neighboring nodes
  o Or another augmentation

* Can we build augmentation for an infinite lattice?
  o See homework exercise (check tomorrow night)
Can we augment other graphs?
  - $G=(V,E)$ (i.e. a lattice) with distance known
  - Random augmentation adds one shortcut per node
  - Is routing on $G$ + shortcuts used incidentally efficient?

Indeed all these graphs are polylog augmentable:
  - Bounded ball growth, Doubling dimensions
  - Bounded “width” (Trees, bounded treewidth graphs)

What about all graphs? Lower Bound $O(n^{1/\sqrt{\ln(n)}})$
Practical follow up

Can we observe harmonic distribution?
- Yes, using closeness rank instead of distance

Can we prove it emerge?
- Recent results
- Through rewiring, mobility

Geographic routing in social networks.
D. Liben-Nowell et. al. PNAS (2005)
Milgram’s experiment prove that social networks are navigable
- individuals can take advantage of short paths
- with basic information
This is at odds with uniform random graphs
The key ingredients to explain navigability
- A space easy to route (e.g. grid, trees, etc.).
- A subtle harmonic augmentation (e.g. ball radius).