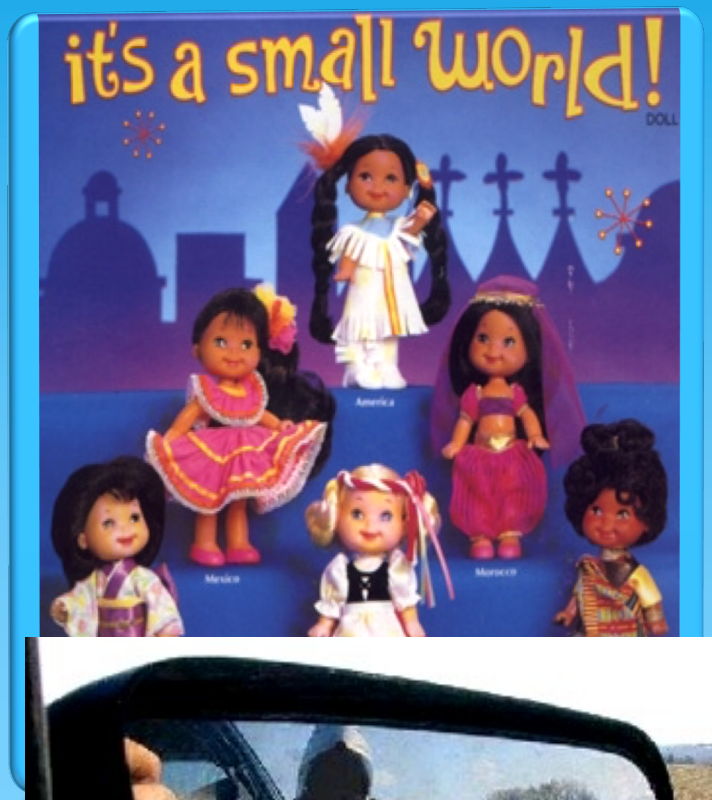
Lecture 1: Structure (1)

Do we live in a combinatorial small world?

COMS 6998: Analysis of Networks & Crowds

Thursday, September 12th

What is the small world effect?



"You're in NY, by the way, do you happen to know x?

- In fact, yes, what a small world!"

What seems remotely distant is indeed socially close.

Why study small world?

* Small world hypothesis is ***necessary*** for
  - Fast and wide information propagation
  - Tipping point: small change has large effect
  - Network robustness and consistency
* Studying the conditions of the "small world" effects tells us a lot about ***how*** we are connected?
* "Man is, by nature, an animal of a society"
  "ζῷον πολιτικὸν"

Small world: a simplistic argument

* Remember how the King lost his fortune to the chess player?
* What would you turn down an offer?
  - An daily doubling series {1¢,2¢,4¢, etc.} over a month
  - Against $1,000? YES/NO
  - Against $100,000? YES/NO
  - Against $1,000,000? YES/NO
  - Against $100,000,000? YES/NO

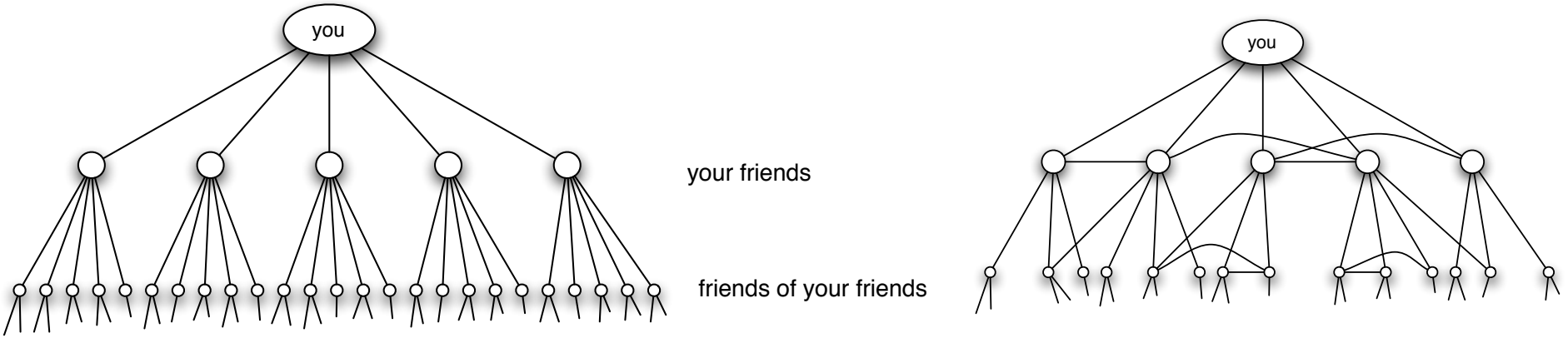
Small world: a simplistic argument

* How many people would you recognize by name?
  * ‘67 M. Gurevitch (MIT): about 500
* Roughly, how many are socially related to you?

	how close to you?	Compares to	%. US pop.
500	direct acquaintance	C.S. dept	0.00017%
250,000	share an acquaintance with you	Harlem district	0.083%
125m	share an acquaintance with a friend of yours	Northeast + Midwest	42%

Small world: The skeptics

* The previous model is way too optimistic!

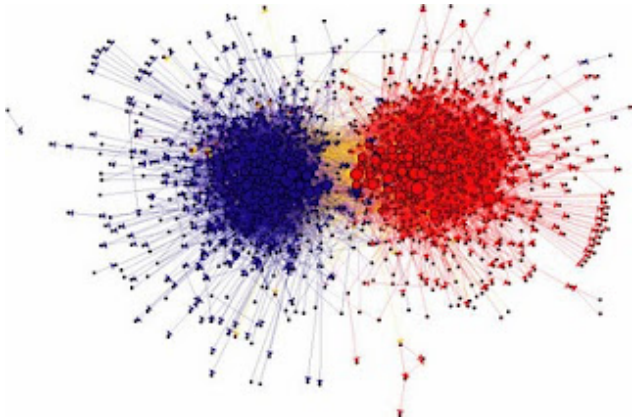


  - Reason #1: it assumes acquaintance set are disjoint
    * whereas they are related as part of a ***social graph***
    * expansion through the graph may be limited.

CS@CU

Small world: The skeptics

* The previous model is way too optimistic!



  - Reason #2: social acquaintances are ***biased***

    Geography, occupation & social status, race

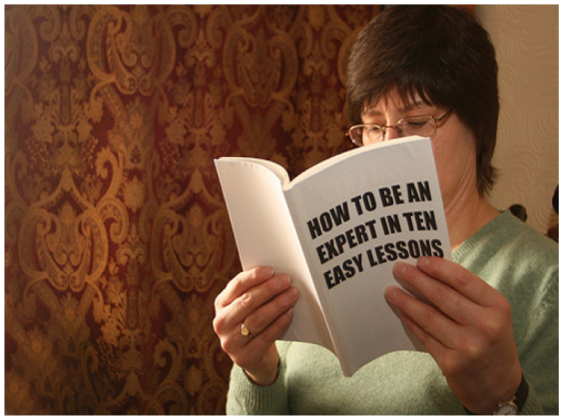
    * Favors clusters, inbreeding. Increases social distance
  - Others alternatives: there are power users

Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
* It's a "complex small world"
* It's an "algorithmic small world"

The 10 papers
that will make you a social expert



10 sociological must-reads

1. S.Milgram, "The small world problem," *Psychology today*, 1967.
2. M. Granovetter, "The strength of weak ties: A network theory revisited," *Sociological theory*, vol. 1, pp. 201–233, 1983.
3. M. McPherson, L. Smith-Lovin, and J. M. Cook, "Birds of a Feather: Homophily in Social Networks," *Annual review of sociology*, vol. 27, pp. 415–444, Jan. 2001.
4. M. O. Lorenz, "Methods of measuring the concentration of wealth," *Publications of the American Statistical Association*, vol. 9, no. 70, pp. 209–219, 1905.
   + H. Simon, "On a Class of Skew Distribution Functions," Biometrika, vol. 42, no. 3, pp. 425–440, 1955.
5. R. I. M. Dunbar, "Coevolution of Neocortical Size, Group-Size and Language in Humans," *Behav Brain Sci*, vol. 16, no. 4, pp. 681–694, 1993.
6. D. Cartwright and F. Harary, "Structural balance: a generalization of Heider's theory.," *Psychological Review*, vol. 63, no. 5, pp. 277–293, 1956.
7. M. Granovetter, "Threshold Models of Collective Behavior," *The American Journal of Sociology*, vol. 83, no. 6, pp. 1420–1443, May 1978.
8. B. Ryan and N. C. Gross, "The diffusion of hybrid seed corn in two Iowa communities," *Rural sociology*, vol. 8, no. 1, pp. 15–24, 1943.
   + S. Asch, "Opinions and social pressure," *Scientific American*, 1955.
9. R. S. Burt, *Structural Holes: The Social Structure of Competition*. Harvard University Press, 1992.
10. F. Galton, "Vox Populi," *Nature*, vol. 75, no. 1949, pp. 450–451, Mar. 1907.

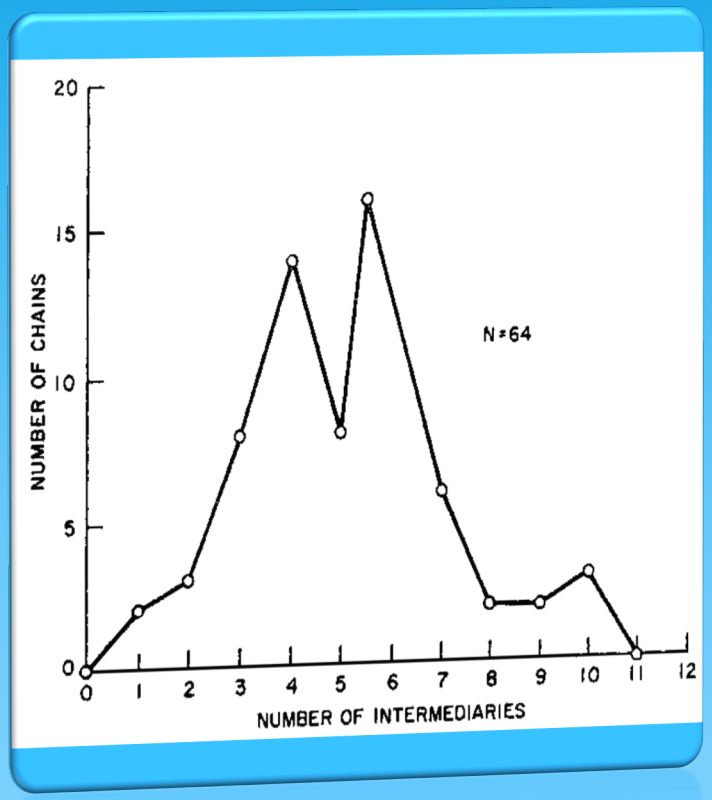
Milgram's experiment

* A direct experimental approach to small world
  - Main issue: we do not know the graph (This is 1967!)
  - Can we still find a method to exhibit small chains of acquaintance between two arbitrary people?
* At least we can try it out:
  * Pick a single "target" and an arbitrary sample of people

  Send a "folder" with target description, asking participants to

  * Add her name, and forward to her friend or acquaintance who she thought would be likely to know the target,
  * Send a "tracer" card directly to a collection point

Would this work?



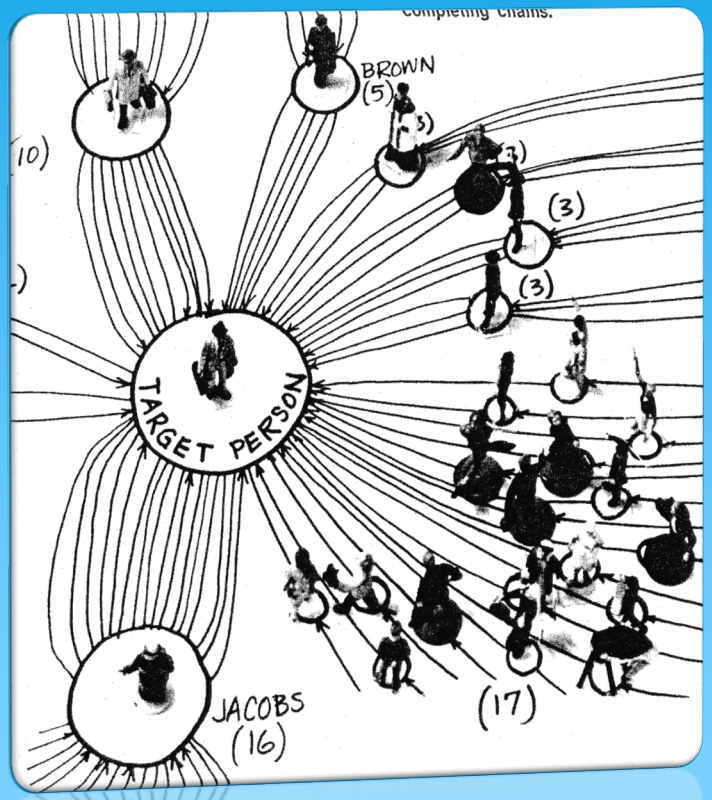
First folders arrived using 4 days and 2 intermediaries

2 successful experiments

- 1st: 44 folders out of 160
- 2nd : 64 folders on 296
- Average distance = 6.2.

The small world problem, S. Milgram, *Psychology today* (1967)

More observations



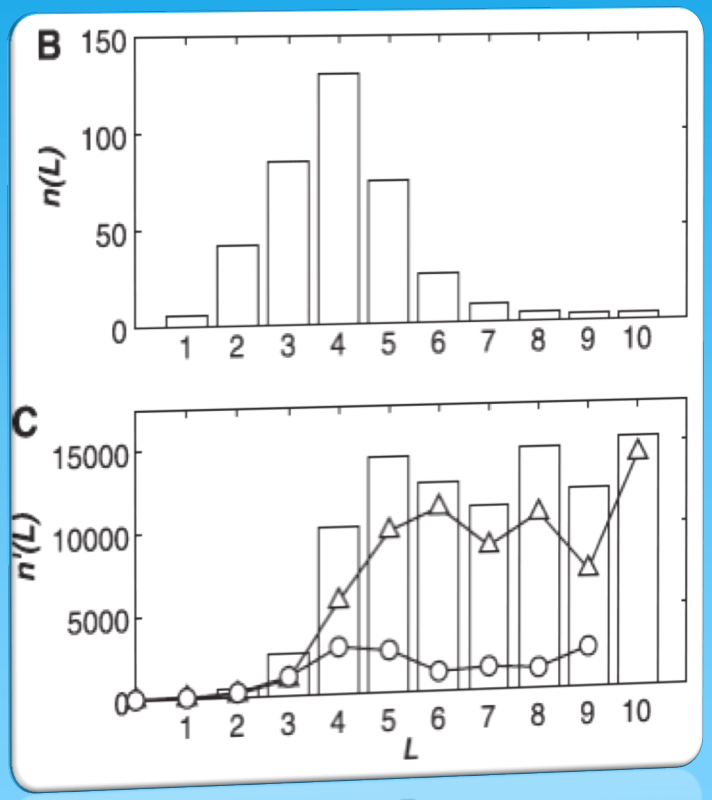
- Chains not random
  locally: gender, family/ friends
  globally: hometown/job
- Some received multiple folders (up to 16)
- Drop out-rate ~25%
  May artificially reduce distance and can be corrected

The small world problem, S. Milgram, *Psychology today* (1967)

Experiment sequels

* Are these results reproducible?
  * Yes, using new forms of communication
    P. Dodds, R. Muhamad, and D. Watts. An experimental study of search in global social networks. *Science* (2003).
* Another method: assuming graph is known
  * Exhaustive search:
  * J. Leskovec and E. Horvitz. Planetary-scale views on a large instant-messaging network. WWW (2008)
  * Erdös number, Kevin Bacon number
  * More evidence of short paths, including in different domains

Confirmation using email

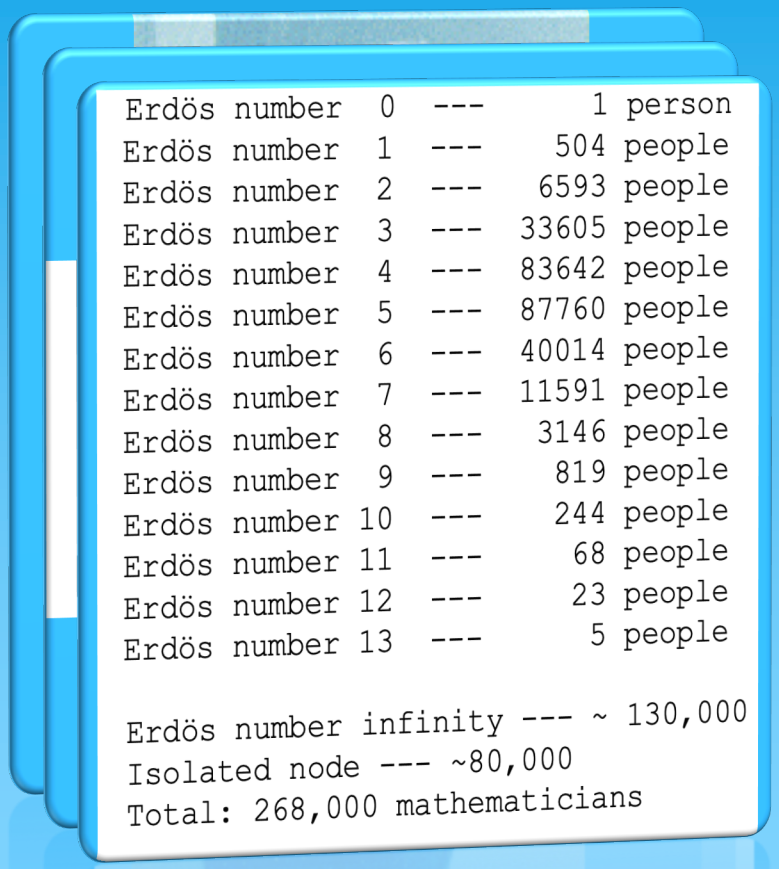


- Larger scale
  18 targets, 24,000 starting points
- Drop-out rates : 65%
  success rate is lower: 1.5%
  median distance:
  4 raw data and 7 corrected data
- Role of social status
  success rate changes, but not mean distance

Even today: Sign-up at smallworld.sandbox.yahoo.com

An experimental study of search in global social networks.
P. Dodds, R. Muhamad, and D. Watts. *Science* (2003).

Erdös number



Connections = collaboration

- 2 persons connected if they co-authored a paper.

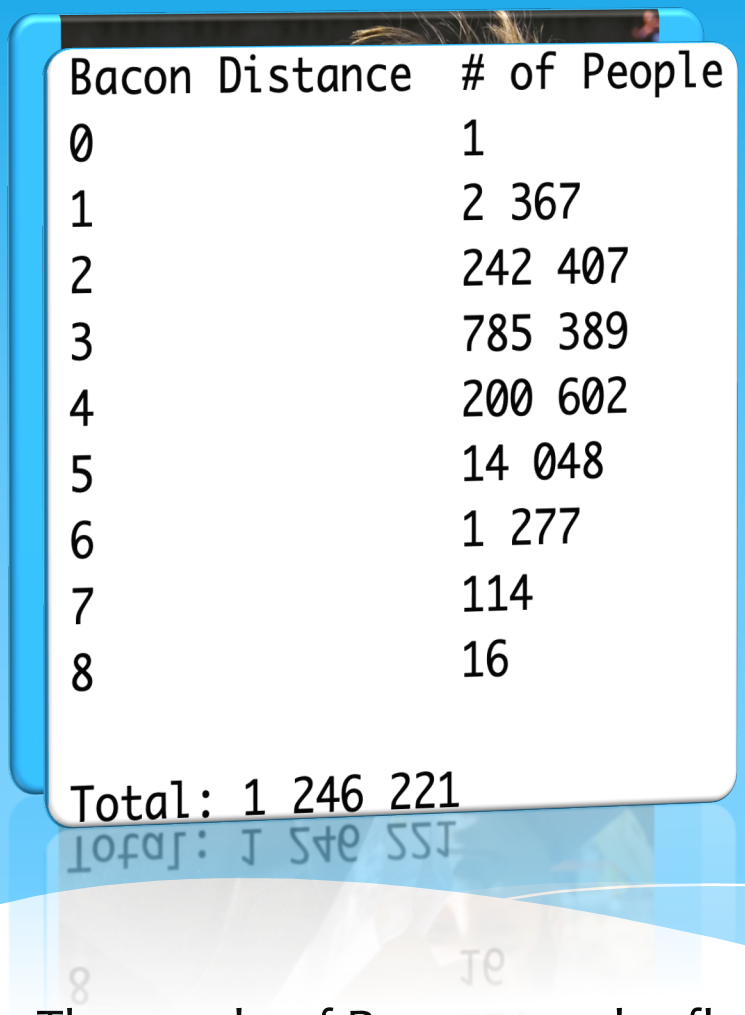
How far are you from Erdös?

- 94% within distance 6
- Median: 5
- Small number prestigious

  Field medalists: 2-5; Abel prize: 2-4

The Erdös Number Project, www.oakland.edu/enp/

Kevin Bacon number



Bacon Distance	# of People
0	1
1	2 367
2	242 407
3	785 389
4	200 602
5	14 048
6	1 277
7	114
8	16

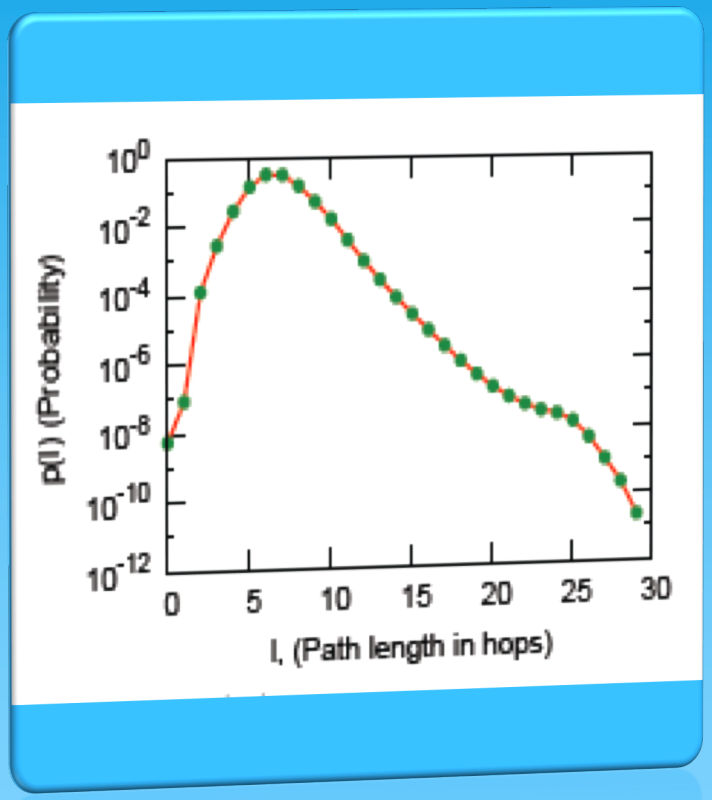
Total: 1 246 221

Connections = collaboration

- 2 actors connected if they appear in one movie
- ~ 88% actors connected
- 98.3% actors: 4 and less
  Maximum number is 8.

The oracle of Bacon, oracleofbacon.org

Instant Messaging



Different scales

- 180m nodes, 13b edges

Connections = collaboration

- 180m nodes, 1.3b edges

Planetary-scale views on a large instant-messaging network.
J. Leskovec and E. Horvitz. WWW (2008)

Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
* It's a "complex small world"
* It's an "algorithmic small world"

Brief Recap

Small-World

* Each of us maintains a set of acquaintances
* Fast (geometric) expansion implies that we reach anyone through a few hops.
* We are all pals

Big-World

* This looks too good
  * Our friends = a graph, limiting expansion
  * We are biased, factors increase distance

Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
  - Model of large random graphs
  - Properties & Thresholds (Probabilistic method)
  - A words on regular graph
* It's a "complex small world"
* Thursday: It's an "algorithmic small world"

Do short paths exist?

* Yes, according to our combinatorial argument shown before (assumes disjoint sets, simplistic)
  - Can we extend it to more ***realistic*** model?
* Approach #1:
  1. construct a graph from model of social connections
  2. Characterizes connectivity, distance, diameter
  - Main difficulty: connections between people are difficult to predict

Random Graphs

* Uniform random graphs (Erdös-Rényi), 3 flavors
  1. Each edge exists independently (with fixed prob. p)
  2. Choose a subset of m edges
  3. Choose for each node d random neighbors
* Main merit: properties are established rigourously for large scale asymptotic (size n going to ∞)
  - Properties of connected components
  - Presence/absence of isolated nodes
  - Small diameter

Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
  - Model of large random graphs
  - Properties & Thresholds (Probabilistic method)
  - A words on regular graph
* It's a "complex small world"
* Thursday: It's an "algorithmic small world"

How to look at large graphs?

* Consider increasing sequence of graphs:
  - $G_n=(V_n,E_n)$, where $V_n=\{1,2,\ldots,n\}$
  - E_n is random (e.g., edge {i,j} occurs w.p. p(n))

* When is an assertion A satisfied?
  - Find a threshold function: a function t(n) such that

$$(i)\ \text{When } \lim_{n\to\infty}\frac{p(n)}{t(n)}=0,\ \text{then } \mathbb{P}\left[G_n \ \text{ satisfies } \ A\right]\to_{n\to\infty} 0.$$

$$(ii)\ \text{When } \lim_{n\to\infty}\frac{p(n)}{t(n)}=\infty,\ \text{then } \mathbb{P}\left[G_n \ \text{ satisfies } \ A\right]\to_{n\to\infty} 1.$$

Example #1: first edge

* Let A be the property of having at least 1 edge
  * Seems like the minimum to ask for a graph
* THM 1: $t(n) = 1/n^2$ is a threshold for having 1 edge

$$\begin{aligned}
&\mathbb{P}(\text{Graph is Empty}) \\
&= (1 - p(n))^{\frac{n(n-1)}{2}} \\
&= \exp\left(\frac{n(n-1)}{2} \cdot \ln(1 - p(n))\right) \\
&= \Theta(\exp\left(-n^2 p(n)\right))
\end{aligned}$$

Example #2: 3 connected nodes

* Let A be {G_n has at least 3 connected nodes}
  * i.e., G_n satisfies A if has a pair of adjacent edges
* P[G_n satisfies A] not easy to write (i.e. matching)
* An example of the probabilistic method
  * Let N_n=the number of adjacent edges in G_n
  * We have {G_n satisfies A} = {$N_n > 0$}
  * And a special way to write N_n $$N_n = \sum_{e,e' \text{ adj.}} X_{e,e'}$$

Example #2: Average Analysis

* Let A be {G_n has at least 3 connected nodes}
* Let N_n be #{adjacent edges} $N_n = \sum_{e,e' \text{ adj.}} X_{e,e'}$
  - Let's compute E[N_n]
  - First, by linearity of expectation

$$E[N_n] = E\left[\sum_{e,e' \text{adj.}} X_{e,e'}\right] = \sum_{e,e' \text{adj.}} E\left[X_{e,e'}\right] = \sum_{e,e' \text{adj.}} p(n)^2$$

CS@CU

Example #2: Average Analysis

* Let A be {G_n has at least 3 connected nodes}
* Let N_n be #{adjacent edges} $N_n = \sum_{e,e' \text{ adj.}} X_{e,e'}$
  - Let's compute E[N_n]
  - Second, counting number of pairs of adjacent edges

$$E[N_n] = \sum_{e,e' \text{adj.}} p(n)^2 = \binom{n}{2} \cdot 2 \cdot (n-2) \cdot \frac{1}{2} \cdot p(n)^2 = \frac{n(n-1)(n-2)}{2} p(n)^2 = \Theta\left(\left(n^{3/2} \cdot p(n)\right)^2\right)$$

CS @CU

Example #2: Average Analysis

* Let A be {G_n has at least 3 connected nodes}
  * The threshold seems to be $p(n) = 1/n^{3/2}$, let's prove it
* Let N_n be #{adjacent edges}

$$N_n = \sum_{e,e' \text{ adj.}} X_{e,e'}$$

  * What we have shown

$$\frac{p(n)}{1/n^{3/2}} \to_{n\to\infty} 0 \implies E[N_n] \to_{n\to\infty} 0$$

$$\frac{p(n)}{1/n^{3/2}} \to_{n\to\infty} \infty \implies E[N_n] \to_{n\to\infty} \infty$$

  * What we would like to show

$$\frac{p(n)}{1/n^{3/2}} \to_{n\to\infty} 0 \implies P[N_n > 0] \to_{n\to\infty} 0$$

$$\frac{p(n)}{1/n^{3/2}} \to_{n\to\infty} \infty \implies P[N_n > 0] \to_{n\to\infty} 1 \quad (i.e., P[N_n = 0] \to 0)$$

CS @CU

Tool #1: Markov Inequality

* THM: For any X non-negative, and a real number *a*

$$\mathbb{P}\left[X \geq a\right] \leq \frac{\mathbb{E}\left[X\right]}{a}.$$

* Can we deduce that no adjacent edges exist?
  - Yes, because we characterize E[N_n]

$$P[N_n > 0] = P[N_n \geq 1] \leq \frac{E[N_n]}{1}$$

  - Hence $E[N_n] \to 0 \implies P[N_n > 0] \to 0$

Tool #2: Concentration inequality

* We wish to apply the same proof in the other sense
  - Unfortunately, $E[N_n] \to \infty \not\Rightarrow P[N_n > 0] \to \infty$
    We need a concentration result (i.e., X close to E[X])

* THM: $P[X = 0] \leq \frac{E\left[(X - E[X])^2\right]}{E[X]^2} \leq \frac{Var(X)}{E[X]^2}$

$$\text{Let } Y = (X - E[X])^2$$

$$E[Y] = Var(X)$$

$$P[X = 0] \leq P[Y \geq E[X]^2] \leq \frac{E[Y]}{E[X]^2}$$

CS @CU

Tool #2: Another inequality

* THM: $P[X = 0] \leq \frac{Var(X)}{E[X]^2}$
  * Great, but how can we bound Var(X)?
* If X is a sum of variables, then we can use
  * Lemma: Assume $X = \sum_{i \in \mathcal{I}} S_i, \quad \text{where } \forall i \in \mathcal{I}, 0 \leq S_i \leq 1$
    * Then $\text{Var}[X] \leq E[X] + \sum_{i \neq j} \text{Cov}(S_i, S_j)$
  * Where $\text{Cov}(S, S') = E\left[(S - E[S]) \cdot (S' - E[S'])\right]$
  * Note that, S and S' independent implies Cov(S,S')=0

Finishing the proof

* So $P[N_n = 0] \leq \frac{Var[N_n]}{E[N_n]^2} \leq \frac{E[N_n] + \sum_{e,e' \text{adj.} \neq f,f' \text{adj.}} \text{Cov}(X_{e,e'}, X_{f,f'})}{E[N_n]^2}$

* What is $Cov(X_{e,e'}, X_{f,f'})$?
  * If $|\{e, e'\} \cap \{f, f'\}| = 0$ then covariance is 0
  * We can't have $|\{e, e'\} \cap \{f, f'\}| = 2$ they are distinct
  * What about $|\{e, e'\} \cap \{f, f'\}| = 1$?
    * In that case $\text{Cov}(X_{e,e'}, X_{f,f'}) \leq E[X_{e,e'} X_{f,f'}] = p(n)^3$

CS@CU

Finishing the proof (2)

* How many terms with $|\{e, e'\} \cap \{f, f'\}| = 1$?
  - Need to choose
  - the set {e,e'} n (n-1) (n-2) / 2 choices
  - Wheter edge f=e or f=e' 2 choices
  - The edge f' adjacent to f (n-3)
  - Total n (n-1) (n-2) (n-3), approximately n^4 such terms

Finishing the proof (3)

* So $P[N_n = 0] \leq \frac{Var[N_n]}{E[N_n]^2} \leq \frac{E[N_n] + \sum_{e,e' \text{adj.} \neq f,f' \text{adj.}} \text{Cov}(X_{e,e'}, X_{f,f'})}{E[N_n]^2}$
* Remember $E[N_n] = \Theta\left((n^{3/2}p(n))^2\right)$
* Hence $P[N_n = 0] \leq \text{cst} \cdot \frac{(n^{3/2}p(n))^2 + n^4 p(n)^3}{(n^{3/2}p(n))^2)^2}$

$$\leq \text{cst} \cdot \frac{1}{n^{3/2}p(n)^2} + \frac{1}{n^2 p(n)}$$

  - $\rightarrow_{n\rightarrow\infty} 0$ as $n^{3/2} \cdot p(n) \rightarrow \infty$

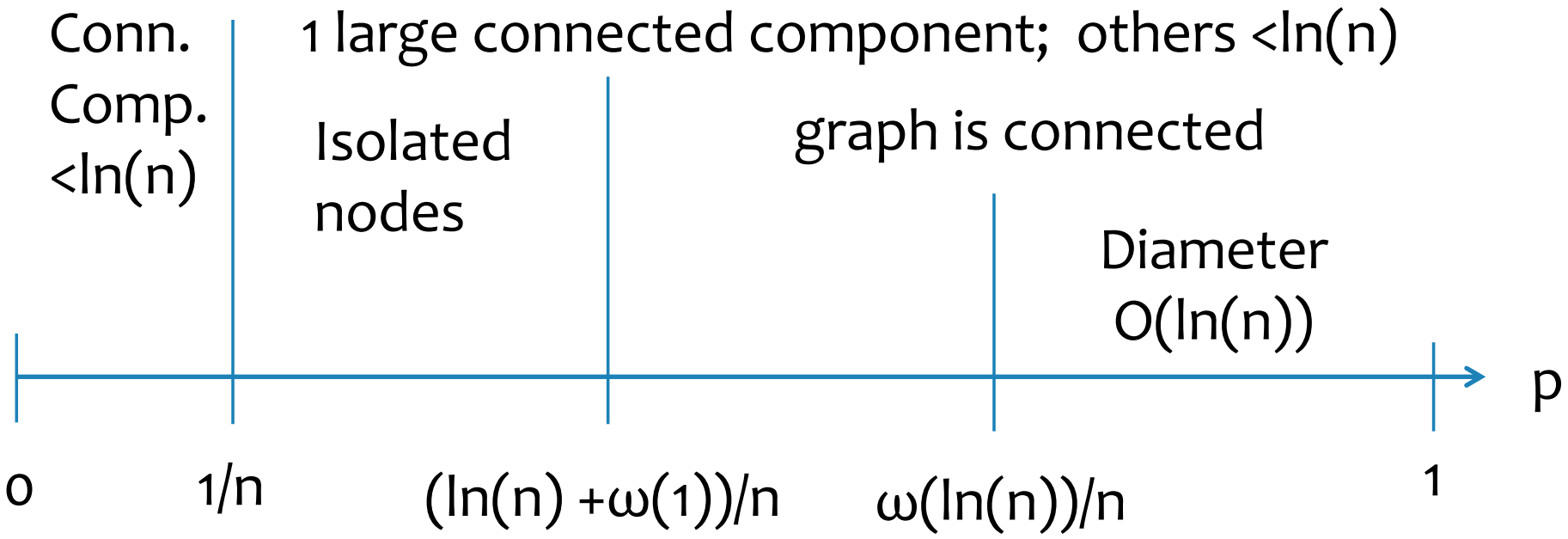
Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
  - Model of large random graphs
  - Properties & Thresholds (Probabilistic method)
  - A words on regular graph
* It's a "complex small world"
* Thursday: It's an "algorithmic small world"

Properties of Unif. Rand. Graph

* Flavor 1 and 2 : thresholds (aka phase transitions)



* All monotone properties have sharp threshold
  - "if G satisfies P, any graph with more edges does"

E. Friedgut, G. Kalai, Proc. AMS (1996) following Erdös-Rényi's results

CS @CU

Outline

* Milgram's "small world" experiment

* It's a "combinatorial small world"
  - Model of large random graphs
  - Properties & Thresholds (Probabilistic method)
  - A words on regular graph
* It's a "complex small world"
* Thursday: It's an "algorithmic small world"

Properties of Unif. Rand. Graph

* Flavor 3 (each node with d random neighbors)
* How to construct?
  1. Assign d "semi-edge" to each node
  2. Pair "semi-edge" into edge randomly
  3. Remove self-loops and multi-edges (rejection)

  (NB: This construction can be used for any degree dist.)

* Degree constant: should we expect connectivity?

Properties of Unif. Rand. Graph

* Transitions is even faster:
  - d=1: collection of disjoint pairs (not connected)
  - d=2: collection of disjoint cycles (a.s. disconnected)
  - d=3: connected, small diameter, in fact much more ...
* Thm: for d≥3 there exists γ>0 such that for any size G is a γ_expander with high probability:

$$\text{For any subset } A \subseteq V, \quad \frac{|\partial(A, A^c)|}{\min(|A|, |A^c|)} \geq \gamma .$$

$$\partial(A, B) = \{ (u, v) \in E \mid u \in A \text{ and } v \in B \}$$

CS@CU

Properties of expanders

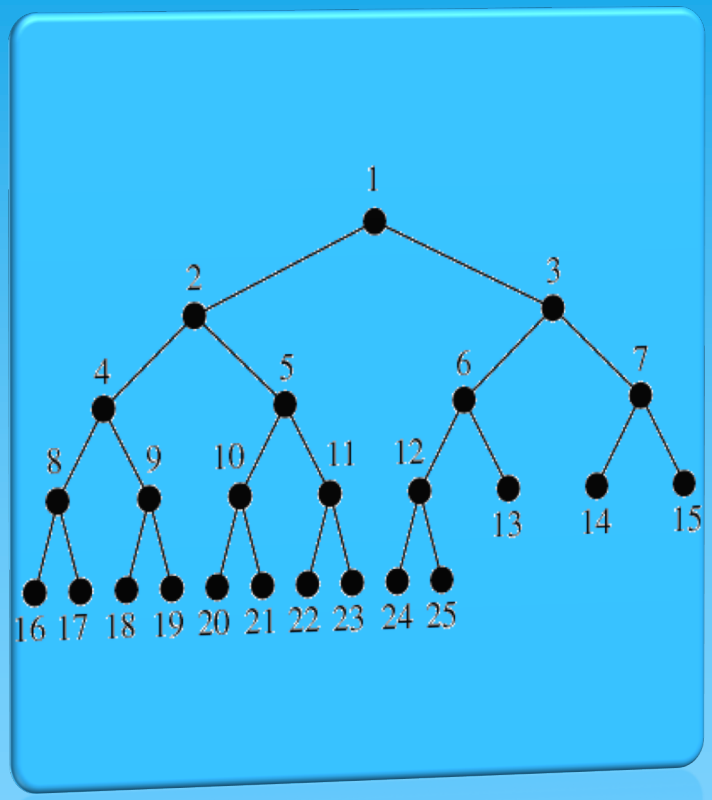
* Proposition 1: If G is a γ_expander with degree at most d, its diameter is at most $\frac{2d}{\gamma}\ln(n)+1$
  * Proof: combinatorial geometric expansion

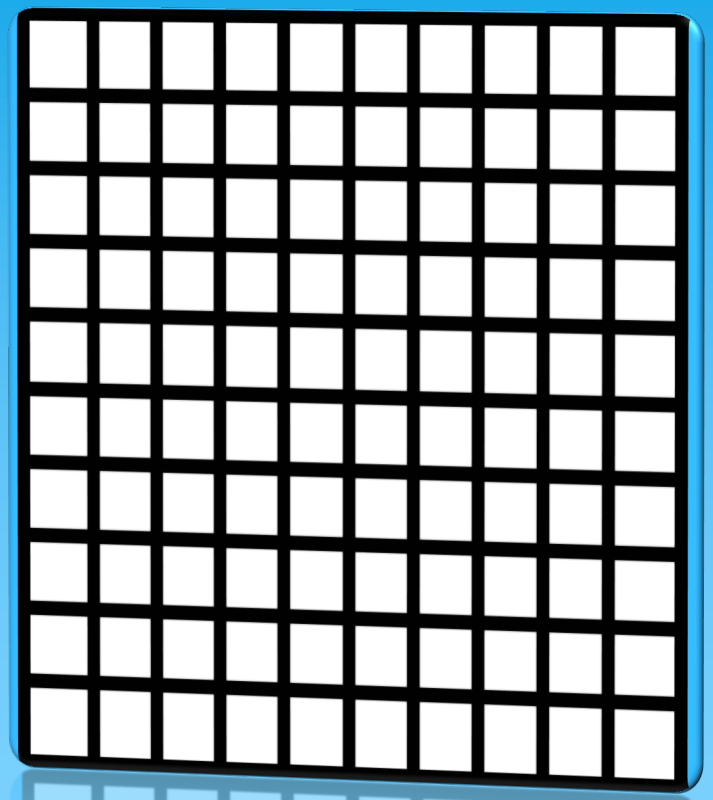
$$S_0 = \{s\} \text{ and } S_{j+1} = S_j \cup N(S_j)$$

$$N(A) = \{\ v \in V \mid v \notin A, \text{ there exists } u \in A \text{ such that } (u,v) \in E\ \}$$

* Expanders have other properties
  * Robustness w.r.t. edges removal
  * Behavior of random walks

Are these graphs expanders?





For any subset $A \subseteq V$, $\dfrac{|\partial(A, A^c)|}{\min(|A|, |A^c|)} \geq \gamma$.

$$\partial(A, B) = \{\ (u, v) \in E \mid u \in A \text{ and } v \in B\ \}$$

Summary

* Assuming random uniform connections
  - Phase transitions (i.e. tipping points) are the rule.
  - Connectivity, small diameter, even expander property can be guaranteed without strict coordination
    … this may outclass deterministic structure!
  - Elegant and precise mathematical formulation

* Is small world only a reflection of the combinatorial power of randomness?