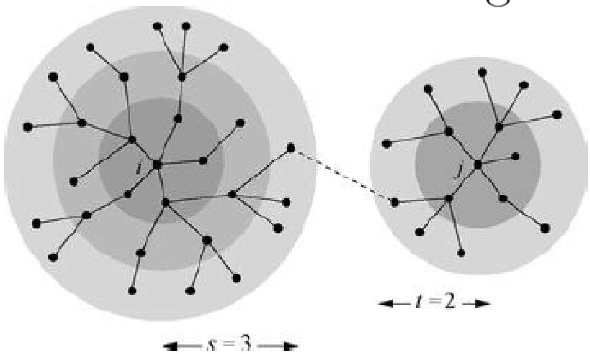
- path lengths
  - Intuition: avg number of nodes s steps away from random i is c^s. We reach all vertices when $c^s \approx n$, i.e., $s \approx \ln n / \ln c$.
  - Problem: this argument doesn't work when s is large.

  - Consider two starting vertices i and j with their s- and $t-$distance neighborhoods, respectively, when s, t are small



1. if "- - - - -" exists between surfaces then one can show that there are edges between larger surfaces

$$\Longrightarrow \quad \Pr[d_{ij} > s+t+1] \approx \text{prob } \nexists \text{ edge between two surfaces}$$

c^s, and c^t when t is small

2. There are on avg $c^s \times c^t$ pairs of nodes, s.t. one lies on each surface and each pair is connected with prob $p = c/(n-1)$
   i.e., $\Pr[d_{ij} > s+t+1] = (1-p)^{c^{s+t}} = (1-c/n)^{c^{l-1}}$ or $\ln \Pr[d_{ij} > l] = c^{l-1} \ln(1-c/n) \approx -c^l/n$

   l = s+t+1

Networks	# of nodes	Diameter					Modularity				
		Observed	Expected	% difference	Z-score[14]	P-value	Observed	Expected	% difference	Z-score[14]	P-value
Characters in "Les Miserables"[1]	77	2.64	2.50	5.6	3.58	0.0003	0.56	0.29	93.4	30.12	$<10^{-4}$
Words in "David Copperfield"[2]	112	2.54	2.48	2.3	1.81	0.0703	0.31	0.29	4.8	1.67	0.0949
Dolphins[3]	62	3.36	2.70	24.3	14.40	$<10^{-4}$	0.53	0.37	40.8	11.59	$<10^{-4}$
Political blogs[4]	1224	2.74	2.59	5.7	23.5	$<10^{-4}$	0.43	0.14	206.9	189.27	$<10^{-4}$
Co-authorship[5]	7610	7.03	5.42	29.6	64.70	$<10^{-4}$	0.81	0.49	64.9	12.50	$<10^{-4}$
Football[6]	115	2.51	2.23	12.5	54.30	$<10^{-4}$	0.60	0.28	119.2	44.68	$<10^{-4}$
Power[7]	4941	18.99	8.32	128.3	14.30	$<10^{-4}$	0.93	0.73	28.5	105.10	$<10^{-4}$
Airline[8]	810	3.06	2.61	17.4	3.53	0.0004	0.31	0.13	130.0	114.70	$<10^{-4}$
Electronic circuits[9]	512	6.86	5.64	21.6	12.40	$<10^{-4}$	0.81	0.63	28.6	35.96	$<10^{-4}$
Protein-protein interaction[10]	1870	6.81	5.78	17.8	9.19	$<10^{-4}$	0.81	0.72	13.2	18.23	$<10^{-4}$
Neural[11]	297	2.46	2.35	4.5	3.38	0.0007	0.40	0.22	80.0	51.26	$<10^{-4}$
Transcriptional regulatory[12]	3459	3.72	3.39	9.7	3.60	0.0003	0.60	0.47	29.5	58.29	$<10^{-4}$
Metabolic[13]	563	8.78	6.54	34.3	18.67	$<10^{-4}$	0.84	0.73	14.5	14.72	$<10^{-4}$

[1]The network of coappearances of characters in Victor Hugo's novel "Les Miserables". Nodes represent characters and edges connect any pair of characters that appear in the same chapter.
[2]The network of common adjective and noun adjacencies for the novel "David Copperfield" by Charles Dickens. Nodes represent the most commonly occurring adjectives and nouns in the book.
[3]The network of frequent associations between 62 dolphins in a community living off Doubtful Sound, New Zealand.
[4]The network of political blogs. Nodes represent blogs and edges are the links between blogs.
[5]The network of scientists posting preprints on the high-energy theory archive at www.arxiv.org, 1995–1999. Nodes are authors and edges connect coauthors.
[6]The network of American football games between Division IA colleges during regular season Fall 2000. Nodes are teams and edges connect teams that contest in a game.
[7]The network of the Western States Power Grid of the United States. Nodes are power plants, stations and households, and edges are powerlines.
[8]The network of scheduled air line connections in United States, 2005. Nodes are airports and edges are scheduled direct flights.
[9]Electronic circuits. Nodes are electronic elements and edges are electronic connections.
[10]The protein-protein interaction network of the budding yeast *S. cerevisiae*. Nodes are proteins and edges connect proteins that interact with each other.
[11]The neural network for the worm *C. elegans*. Nodes are neurons and edges link neurons that connect.
[12]The transcriptional regulatory network of the budding yeast *S. cerevisiae*. Nodes are genes and edges connect genes that regulate one another.
[13]The metabolic network of the bacterium *E. coli*. Nodes are metabolites and edges connect metabolites that can be converted by a biochemical reaction.
[14]Z-score, number of standard deviations by which the observation deviates from the expectation.

doi:10.1371/journal.pone.0005686.t001

http://complexnt.blogspot.com

Generating Functions and Degree Distributions

The *generating function* (gf) for the probability distribution p_k is the polynomial

$$g(z) = \sum_{k=0}^{\infty} p_k z^k.$$

If we know gf for p_k then we can recover the values of p_k by differentiating

$$p_k = \frac{1}{k!} \frac{d^k g}{dz^k}\bigg|_{z=0}$$

Example: $k = 0, 1, 2$ with the respective $p_k = \frac{1}{2}, \frac{7}{16}, \frac{1}{16}$ for all k then

$$g(z) = \frac{1}{2} + \frac{7}{16} z + \frac{1}{16} z^2$$

Example: k follows Poisson distribution, i.e., $p_k = e^{-c} \frac{c^k}{k!}$

$$g(z) = e^{-c} \sum_{k=0}^{\infty} \frac{c^k}{k!} z^k = e^{c(z-1)}$$

Power-law distributions $p_k = Ck^{-\alpha},\ \alpha > 0,\ k > 0$

Reminder: C is calculated from normalization condition, i.e., $C = 1/\zeta(\alpha)$

$$p_k = \begin{cases} 0 & k = 0 \\ k^{-\alpha}/\zeta(\alpha) & k > 0 \end{cases} \implies g(z) = \frac{1}{\zeta(\alpha)} \sum_{k=1}^{\infty} k^{-\alpha} z^k = \frac{Li_\alpha(z)}{\zeta(\alpha)}$$

Polylogarithm

Since we are interested in differentiating $g(z)$ note that

$$\frac{\partial Li_\alpha(z)}{\partial z} = \frac{Li_{\alpha-1}(z)}{z}$$

Some properties of $g(z)$

- $g(1) = 1$
- $\langle k \rangle = g'(1)$, $\langle k^2 \rangle = \left[\left(z\frac{d}{dz}\right)^2 g(z) \right]_{z=1}$, ... , $\langle k^m \rangle = \left[\left(z\frac{d}{dz}\right)^m g(z) \right]_{z=1}$
- Choose m integers k_i from $p_k \Rightarrow \Pr[\text{chosing particular set of values } \{k_i\}] = \prod_i p_{k_i}$

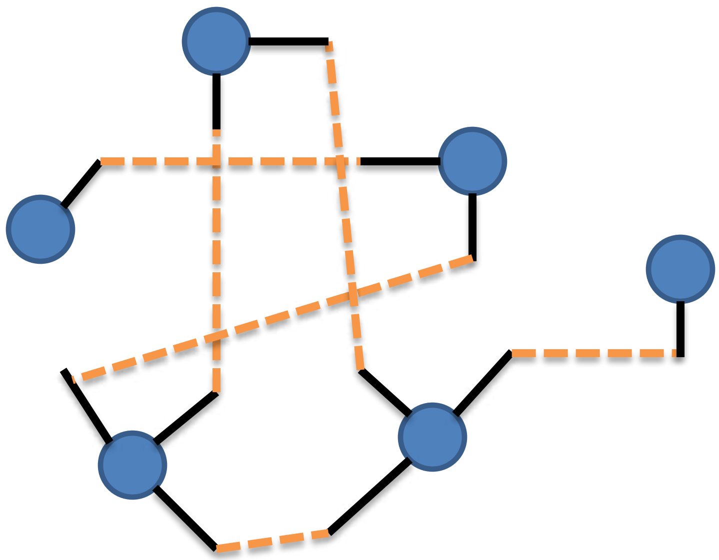
$$\pi_s = \Pr[\sum_{i=1}^{m} k_i = s] = \sum_{k_1=0}^{\infty} \cdots \sum_{k_m=0}^{\infty} \delta(s, \sum_i k_i) \prod_{i=0}^{m} p_{k_i} \Rightarrow$$

$$h(z) = \sum_{s=0}^{\infty} \pi_s z^s = \cdots = \left(\sum_{k=0}^{\infty} p_k z_k \right)^m = (g(z))^m$$

drawn values add to a specific sum *s*

Random Graphs and Configuration Model

Degrees: 1, 1, 2, 2, 3, 3



1. Add *n* nodes
2. Add initial *d(i)* stubs to each *i*
3. Connect stubs iteratively

Problems? Total degree is even; Can create self-loops, multi-edges

Configuration Model

Multi-edges: Probability of adding an edge between i and j with degrees k_i, and k_j is

$$p_{ij} = \frac{k_i k_j}{2m-1}$$

in the limit we can omit -1

Probability of second edge is $(k_i - 1)(k_j - 1)/2m$

Expected number of multiedges in conf model

$$\frac{1}{2(2m)^2}\sum_{ij} k_i k_j (k_i - 1)(k_j - 1) = \frac{1}{2\langle k\rangle^2 n^2}\sum_i k_i(k_i - 1)\sum_j k_j(k_j - 1) = \frac{1}{2}\left[\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}\right]^2$$

Similar result for self-edges

$$\sum_i p_{ii} = \sum_i \frac{k_i(k_i - 1)}{4m} = \frac{\langle k^2\rangle - \langle k\rangle}{2\langle k\rangle}$$

Conclusion? Expected number of multi-edges remains constant as network grows.

Expected number of common neighbors

$$n_{ij} = \sum_l \frac{k_i k_l}{2m}\frac{k_j(k_l - 1)}{2m} = \frac{k_i k_j}{2m}\frac{\sum_l k_l(k_l - 1)}{n\langle k\rangle} = p_{ij}\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}$$

i is connected to *l*

j is connected to *l* given *il*

Final Project

- Download full data set
- Choose one primary type, s.t. the number of full records is >10000 (don't choose "criminal trespass", remove "noisy" records with no time and coordinate)
- Build an undirected network with the following rules
  - Records are nodes
  - *Ij* is an edge iff (1) primary_type(i) = primary_type(j); (2) |time(i) – time(j)|<p1; and (3) dist(location(i) – location(j))<p2, where p1 and p2 are parameters
  - for location use longitude and latitude; for distance use Euclidean distance
- Degree distribution: 1) compute and plot (make sure you adjust the scales, and everything is visible); 2) is it similar to Poisson/binomial, exponential, power law or something else? 3) estimate the chance that "high impact" crime will be connected to another "high-impact" crime (use excess degree distribution and/or degree distr of neighbors)
- Importance of nodes: use various centrality indices to model importance of nodes (compute, plot, explain).
- Compute or estimate clustering coefficient of a network (explain)
- Compute modularity.
- Find clusters in the network (e.g., graclus), plot distribution of sizes
- Can you disconnect the network? What is the smallest size of separator? (e.g., metis)
- Visualize a subnetwork for 1 year (or for less than 1 year)