A discussion of 'Statistical Mechanics of Complex Networks' Part I

Albert & Barabási

Review of Modern Physics, Vol. 74, 2002

Introduction

- cover only parts I, II, and IIIa (pages 1-9)
- more questions than answers...

Small World Networks

What is a small-world network?

- “relatively short” path between any two nodes
- "six degrees of separation"
- *distance* $\equiv$ the shortest path between two nodes
- *diameter* of graph $\equiv$ longest distance between any two nodes
- no hard definition, but diameter similar to random graph ($\sim \ln N$)

Clustering Coefficient

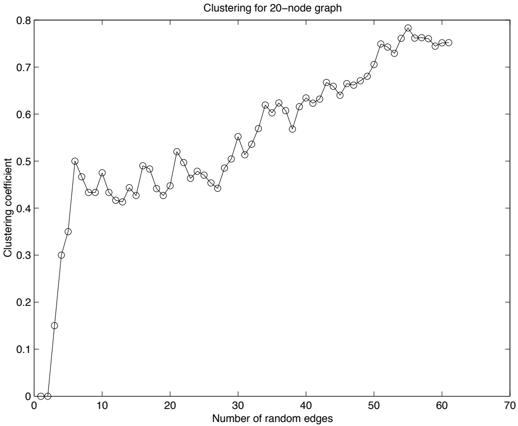
How well do your friends know each other?

- a metric between 0.0 and 1.0
- ratio of edges (E_i) over maximum possible (complete subgraph)
- for each node : $C_i \equiv E_i / \binom{k_i}{2} = \frac{2E_i}{k_i(k_i-1)}$
- for graph, take average over nodes $C = \frac{1}{N} \sum_{i=1}^{N} C_i$

Can be interpreted as

- the probability that two neighbor nodes are connected ($p = C_i$)

Clustering Coefficient Simulations



Clustering Coefficient Observations

Metric is convenient to define, but

- $C = 1$ does NOT imply that graph is completely connected
- $C = 0$ does NOT imply that all nodes are isolated
- C is NOT monotonic as more edges are added (!)

Shortcomings

- doesn't understand the notion of "components"
- uses only one "generation" of information
- "all-or-nothing" metric may be too crude (?)

Is this what we really want to measure . . . ?

Weird Clustering Examples

- a collection of isolated 3-cycles has $C = 1$
- a n-dimensional grid has $C = 0$, although $k = 2n$
- example of graph where adding more edges **lowers** C
  - take two disconnected subgraphs and bridge them

Clustering Examples

Consider a graph with N vertices arranged in

- 1-d ring: k=2, $d(G) = N/2$, $C = 0$ [1]
- 2-d grid: k=4, $d(G) = \sqrt{2N}$, $C = 0$
- 3-d cube: k=6, $d(G) = \sqrt{3}\, N^{1/3}$, $C = 0$
- ...
- n-d hypercube: k=2n, $d(G) = \sqrt{n}\, N^{1/n}$, $C = 0$

Is the last considered a small-world network?

[1] This is why Watts-Strogatz used a 1-d ring with **4** nearest neighbors to bump up C to 3/4.

Scale-Free Networks

Degree distributions

- not all nodes have same degree
- distribution function $P(k)$ denotes probability that random node has k edges
- for random graphs, this is a Poisson distribution with a peak at $\langle k \rangle$ with value $P(\langle k \rangle)$
- for some real networks, the tail of $P(k)$ follows a power-law distribution:
  - $P(k) \sim 1/k^{\gamma} \quad \text{for } 1 < \gamma < 3$
- other real networks exhibit exponential tails
- graphs with $P(k)$ different than Poisson distribution are termed "scale-free"

Issues with Scale-Free Networks

- no hard definition
- why the big fuss? Because physics has properties with power-law tails... (statistical mechanics)
- where does the cut-off for k take effect...?
- the "tail" has tiny portion of nodes... is it really that relevant?

Complex Network Cheat Sheet

	Random Graphs	"Real" Graphs
small-world	YES $d(G) \sim \ln(N)$	YES $d(G) \sim d(G_{random})$
clustering coeff	LOW $(C = p) \ll 0.01$	HIGH ~ 1.0
scale-free	NO Poisson dist.	YES $P(k) \sim 1/k^{\gamma}$ for $1 < \gamma < 3$

Complex Network Cheat Sheet

	Random Graphs	"Real" Graphs	*n*-d lattices
structure	NO	YES (?)	YES
small-world	YES	YES	NO (?)
	$d(G) \sim \ln(N)$	$d(G) \sim d(G_{random})$	$d(G) \sim N^{1/n}$
clustering coeff	LOW	HIGH	0.0
	$(C = p) \ll 0.01$	~ 1.0	$2n$ neighbors
scale-free	NO	YES	YES
	Poisson dist.	$P(k) \sim 1/k^{\gamma}$ for $1 < \gamma < 3$	$P(2n) = 1$, 0 otherwise

Cases Studied

- WWW
- Internet
- movie actors
- science collaboration
- STDs
- cellular networks
- ecological networks
- phone call network
- citation networks
- linguistic networks
- power grid and neural nets
- protein folding

WWW Studies

- at various levels (internet domain, site level, hyperlinks)
- at the hyperlink level:
  - largest network studied (2002)
  - **directed** graph, very **un**symmetric ($k_{out} \ll k_{in}$)
  - both $P_{out}(k)$ and $P_{in}(k)$ how power-law tails
  - with $\gamma_{out} \sim 2.5 \pm 0.25$ and $\gamma_{in} = 2.1$
- Adamic (1999) computed clustering coefficients by making each edge bidirectional (!)
- Faloutsos (1999): an edge is drawn between two domains if there is a least one route that connects them (!)

Movie Actor Collaboration Network

Hooray for IMDb.com!

- Size: half a million actors (!) in 2000
- two actors have an edge if they worked together on a film
- model does not take into account weighted edges (such as # of films worked on together)
- average distance is close to that of random graph

Observations of Table I

- key values:
  - size (N)
  - average degree ($\langle k \rangle$)
  - average distance (ℓ)
- $C \gg C_{rand}$
- $\ell \approx \ell_{rand}$

Things to look at:

- scatter plot of C vs. how "dense" the graph is ($\langle k \rangle / N$)
- scatter plot of density vs. ℓ / ℓ_{rand}

Observations of Table II

- $\gamma_{in} = 2.1$ is quite popular...
- $1 < \gamma < 3$ for both γ_{in} and γ_{out}
- **k** (cut-off) seems pretty high, compared to $\langle k \rangle$
- ℓ_{power} is not as good an estimator as ℓ_{rand} ...

Random Graph Models

Erdös-Rényi (1959)

- N nodes, n edges, chosen randomly from $\binom{N}{2}$ possiblities

Binomial Model

- N nodes, every edge has p probability
- actual # of edges is a random variable
  - Poisson distribution with expected value $p(\binom{N}{2})$
- with $p = n/\binom{N}{2}$ this is **similar** to Erdös-Rényi, but is it the same?

Graph Enumeration

An undirected graph with N vertices,

- has $M \equiv \binom{N}{2} = N(N-1)/2$ possible edges
- # of graphs with exactly n edges, is $\binom{\frac{N(N-1)}{2}}{n}$ = **HUGE!**

Graph Enumeration

An undirected graph with N vertices,

- has $M \equiv \binom{N}{2} = N(N-1)/2$ possible edges
- # of graphs with exactly n edges, is $\binom{\frac{N(N-1)}{2}}{n}$ = **HUGE!**

Given

- $\binom{M}{n} \equiv \frac{M!}{n!(M-n)!}$
- *Stirling's* approximation: $\ln n! \approx n(\ln n - 1)$, for large N

Graph Enumeration Examples

Name	Vertices	$\langle k \rangle$	# graphs
	10	1	3×10^9
	100	1	$\sim 10^{211}$
	1,000	1	$\sim 10^{3,132}$
	10,000	1	$\sim 10^{41,332}$
math authors	70,975	3.9	$\sim 10^{1,200,000}$
movie actors	225,226	61	$\sim 10^{50,000,000}$

If every atom in the universe ($\sim 10^{80}$) was a Petaflop computer, computing since the begining of time (13 billion years ago) you would just need 10^{100} such universes to enumerate the (100, $\langle 1 \rangle$) case...

Do we really know the landscape?

HELP!

- we are looking at only tiny microcosm of graph space for simulations
- how robust are our conclusions?
- importance sampling (?)
- concern with 1-d parameterization ala Watts-Strogatz...
- what if we made random changes to a "real" network? How long before it starts losing its "realness"?
- would any of these metrics help...?

Conclusions and Discussion

- is "small-world" really relevant...? (social networks rarely interact beyond three links...)
- not clear if current metrics really capture the right thing...
  - given $(N, C, \langle k \rangle, \gamma, \ell)$ what can one say about a network?
- introduce new(?) metrics that better recognize components and structure
  - cluster coefficient should be extended for weighted, bidirectional graphs
- power-tail distribution model needs high cut-off values for k
  - what percentage of the available nodes **is** this?

Conclusions ...?

Why is this so hard...?

- because we are trying to **theoritize arbitrary structure** . . .