①

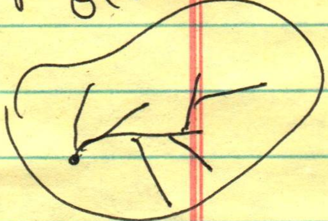
Lecture 12 a – Data incompleteness & sampling
12 b – Models of network growth

Apr 9, 2024
– 11 –

world $G=(V,E)$ $\xrightarrow{g}$ Samples $G'=(V',E')$

Examples :

Dijkstra $O(V^2)$ or $O(V+E\log V)$



WWW ~ 10^{10} vertices

FB (2011) ~ 7.21×10^{8} vertices

$m \sim 10^{10}$

What is the avg path length ?

Reasons to sample

① compute cost
② limited access to data

$\Big\}$ work w/ G'

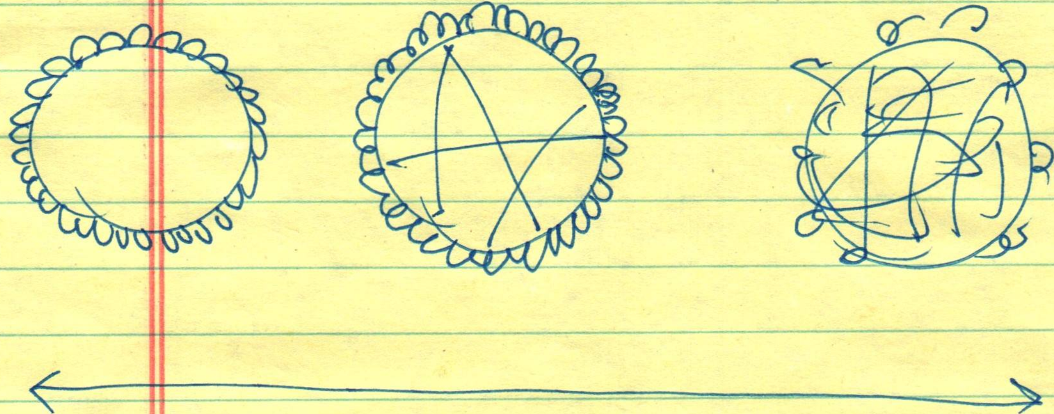
$\Big\{$ $E' \subseteq E$, $V' \subseteq V$

$$f(G) \stackrel{?}{=} f(g(G))$$

interaction

②

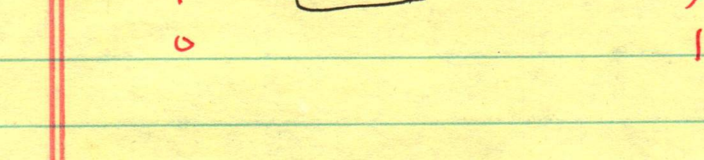
Implications from the Watts-Strogatz model



0 p: prob. rewires some (i, j) 1

C : robust to uniform edge sampling ("local")

$\langle l \rangle$: sensitive to ______ ("global")



3

General approaches to sampling networks.

① Probabilistic sampling (assume access to G ; i.e., full network)

- uniform vertex sampling (gets node i AND all its nbs $j \in nb(i)$)
- uniform edge sampling (incl. each (i,j) w/ prob. p)
- degree-proportional (prob sample $\propto k_i$)
- attribute-prop. (prob. sample $\propto x_i$)
  any local network measure

② Seed-based sampling

- snowball sampling
- BFS ($\Rightarrow$ tree)
- adaptive tree
- random walks

either, for ℓ-step, incl. vtx (& nbs) OR edges

See Lecture Notes

③ Other approaches

- deg sampling : all nodes w/ deg $k_i > k^*$.

4

Sampling induces patterns

>> sparsity
>> biased subgraphs
>> preponderance of low-degree nodes

* Snowball sampling v.s. BFS

* Traceroute ($\approx$ BFS) : an extreme example

  IPv4 , 2^{32} IP addresses .

1

Lecture 12 b : Models of Network Growth

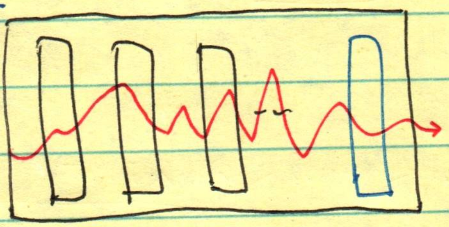
Apr 11, 2024

many networks grow
shrink
dynamically stable

ex: Friendster (2002)
economic exchange ##
(textile industry, 1900s

Temporal network or time-stamped network
(i, j, t) $(i, j, t, \Delta t)$

network representation
(choise to make!)

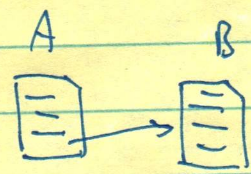


G_t G_{t+1} G_{t+2} ~

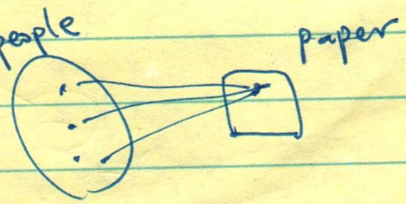
" Another representation Q "
match the snapshot freq. to
the " fundamental " pace of the dynamics

Examples of growing networks :

1) online social nets (FB, IG, TT, www)
2) scientific citation network nodes : papers
(judicial opinions) edges : citations



3) scientific collaboration
nodes : scientists
edges : coauthorships



②

Goal of today: Thinking about "network mechanisms"

model of cause and effect in a network

in terms of edges

Question: where do new edges appear?

$$G_t \rightarrow G_{t+1} = G_t + [V', E']$$

mechanisms → hypotheses (tested; yes or no ?!)

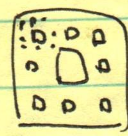
"Scale-free network" (1999)

is it universal?

$$P(k) \propto k^{-\alpha}, \quad \alpha > 1$$

$$\updownarrow$$

$$P(c \cdot k) \propto b \cdot P(k)$$

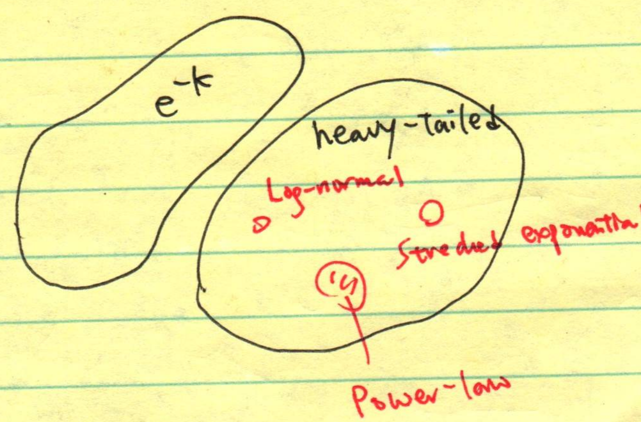


preferential attachment

» network mechanisms for power-law $Pr(k)$.

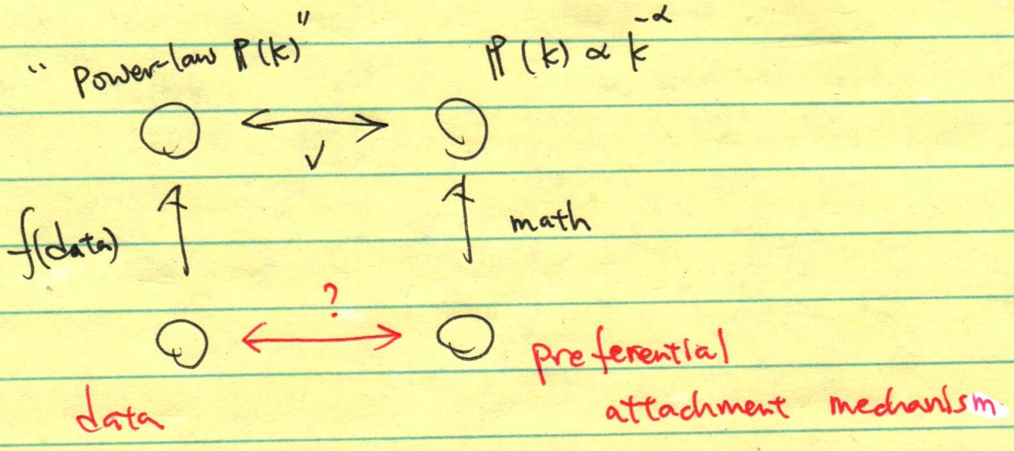
"fractals"

eg. Sierpiński gaskets/triangles

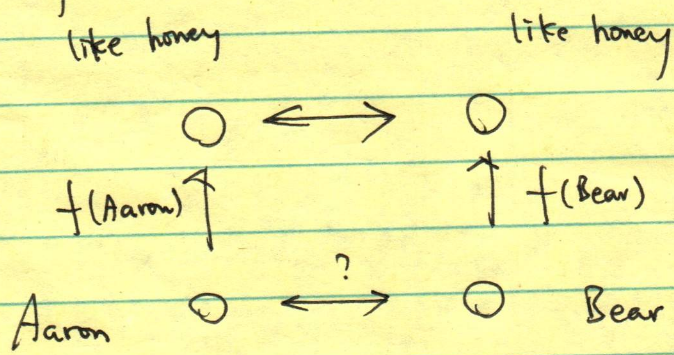


(invented by Aaron)

Honey bear test (late 1990 - early 2000)



Ready to see the honey bear ? ↴



Is Aaron a bear ? What's wrong ?

» measure other things !

" hyperplane that separates classes "

4

Preferential Attachment (invented, thru math Udny Yule, 1925)

"socio grams"

speciation : how many species are in a genus ?

" why is the sizes of the genera so skewed ? "

variation of Polya Urn .

math / combinatorics

math

— Herbert Simon (1950-1960s) Nobel (Ec) / Turing / complex systems

" The Sciences of the Artificial " (book)

exact mathematical form of $\mathbb{P}(k)$

data

— Derek de Solla Price (1960s)

>> citation networks

\+ Simon's model

>> " cumulative advantage "

verbal

— Robert Merton (1960s) sociologist

" Matthew effect "

(rich get richer)

social / financial capital

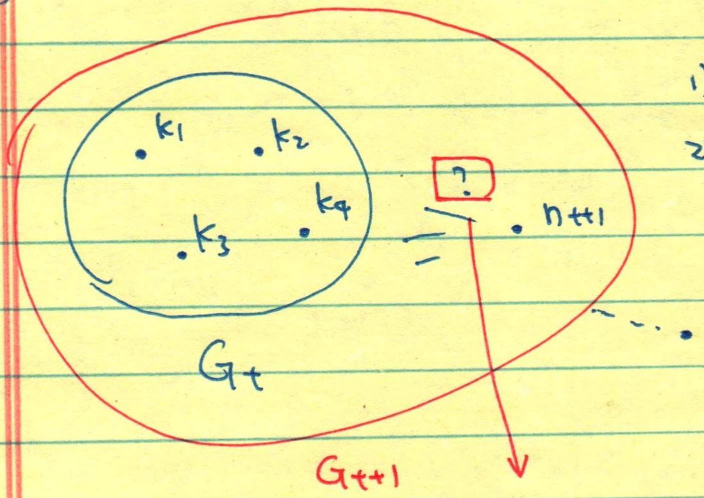
— Albert-László Barabási — Réka Albert (1999)

" Preferential attachment "

lots of work since 1999 (see Lecture Notes)

§ How does PA work?

1) At each time step, add one node.
2) Each new node has degree $c=1$.



"attachment function"

$$P(k_i+1) \propto r + k_i$$

(constant) uniform attachment

preferential part

generalized PA

$$q_i = \left(\frac{r + k_i}{\sum_j (r + k_j)} \right)^{\gamma}$$

PA : $\gamma = 1$

$0 < r < \infty$

$\gamma > 1$ "nonlinear distortion of prob's"

condensation effect.

$\gamma = 2$ (star graph !)

** See Lecture Notes / Textbook for the derivation to the power law.

See Lecture Notes for Price's model viz. (fractal!)

Figure 2

no honey bear fallacy, "zero parameter predictions"
harsh test for the model
1998 - Watt-Strogatz Paper yr=0
2007 - last yr=1

Figure 3 fine-grained evaluation

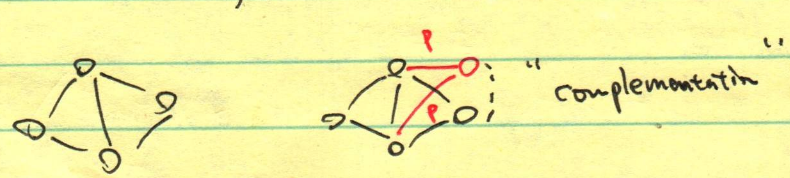
Evidence from
- large-scale (consequences of the model)
- "assumption of linear preferential attachment" validates the input

>> Good causal model!

>> What's missing in the model?
>> just a phenomenological model!

Other models:

Vertex copy mechanism



same scale-free degrees, but abundant Δ's

G_t G_{t+1}

⑦

Finally, see Paper 1: Choosing to grow a graph (WWW'19).

$\boxed{\text{Table 1}}$

"conditional multinomial logit models
(from discrete choice and random utility theory)
unifies existing network growth models"