Social and Information Networks 1: Structure

Internet Analytics (COM-308)

Prof. Matthias Grossglauser
School of Computer and Communication Sciences

EPFL

Overview

- Networks
  - Models for social and technical systems
  - Key properties
- Giant component
  - Random graph model
- Clustering
  - Transitivity
- Strong and weak ties
  - Some ties cluster more than others
- Distances: everything is close
  - Networks are very efficient
- Small world networks
  - Small distances, large clustering not mutually exclusive

Models of networks

- George E. P. Box:
  - "All models are wrong, but some models are useful"
- Many technical and social systems:
  - Networks are central abstractions
- What are models for:
  - (a): Explanation of phenomenon ("physics")
    - Occam's razor: simplest possible model preferred
    - Usually embodies salient features of system
    - Example: economics – skewed income distribution
  - (b): Prediction and inference ("machine learning")
    - Favor performance, not parsimony or explanation
    - Can be "black-box", large # of parameters
    - Example: deep neural networks

Social networks

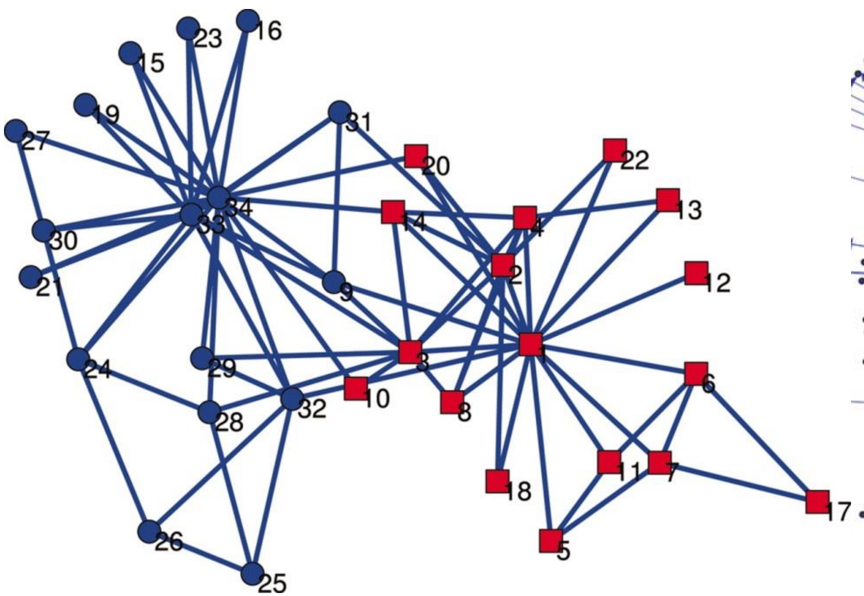
- Ties between individuals and their digital manifestation
- Ties=friendship / business relationship / romantic / shared interest or life situation
- As old as mankind…
  - Or older: many social animals (mammals, insects)
- Social living is complex: explanation for evolution of human intelligence = dealing with large, complex society
  - Trust, behavioral traits, who knows what, trading, rites and rituals,…
  - Dunbar number (~ 150): cognitive limit to group size

Social networks

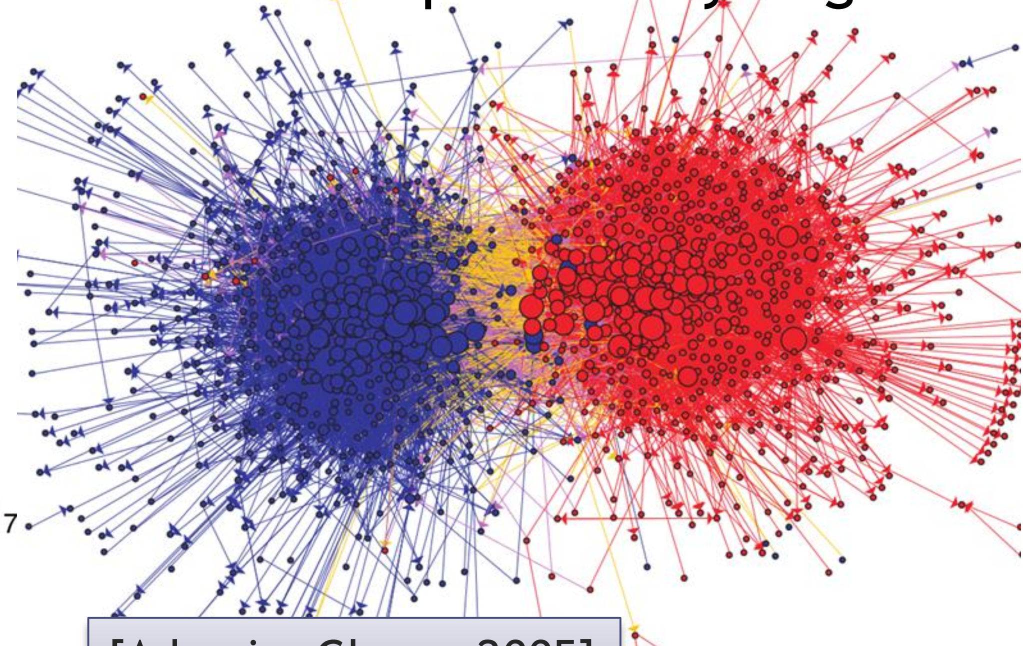
- Tie formation:
  - Complex social process
  - Many types and gradations of ties
- Most parsimonious model: “social network”
  - Graph
  - Undirected edges without attributes or weights
  - Vertices without attributes
  - “Only structure”
- Clearly a strong abstraction!
  - Nevertheless, very rich insights
  - Half a century of research

Social Network Analysis: then and now

- Pre-digital:
  - Sources: surveys etc.
    - Designed, controlled → exactly what you want
  - Scale: small
- Internet:
  - Sources: online databases
    - Research is secondary goal → gaps, inconsistencies, etc.
  - Scale: potentially huge



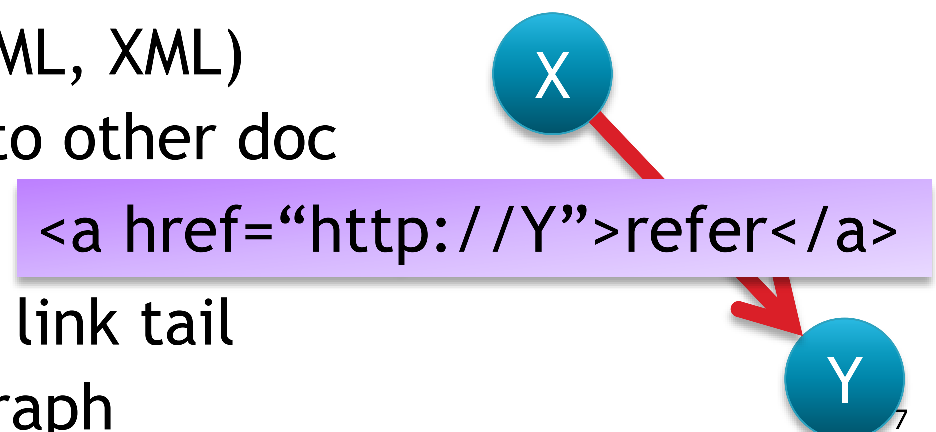
[W. Zachary, 1977]



[Adamic, Glance 2005]

Information networks

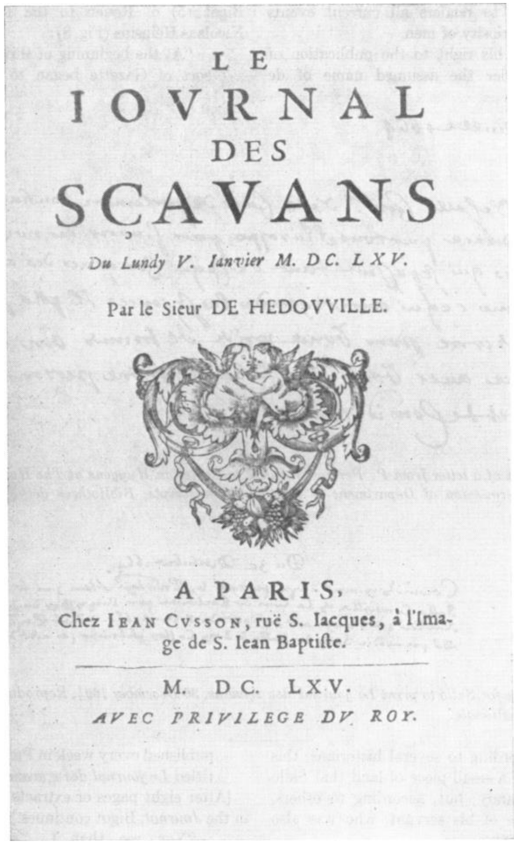
- Web
  - Sir Tim Berners-Lee, CERN, 1989
  - WWW: platform-independent information representation
  - Hypertext concept predates
  - 1990: first HTTP transfer
- Key idea:
  - Universal Resource Locator (URL): global address
  - Structured documents (SGML, XML)
  - Hyperlinks: doc can refer to other doc



- Links are asymmetric
  - Existence under control of link tail
  - Represented as directed graph

Information networks

- Internet
  - Traffic exchange arrangements (peering) w/o central coordination
- Wikipedia
  - References to related concepts
- Scientific literature
  - Bibliography: citing prior related work
- Dictionary
  - Explaining one term in terms of other terms


LE

IOVRNAL

DES

SCAVANS

Du Lundy V. Ianvier M. DC. LXV.

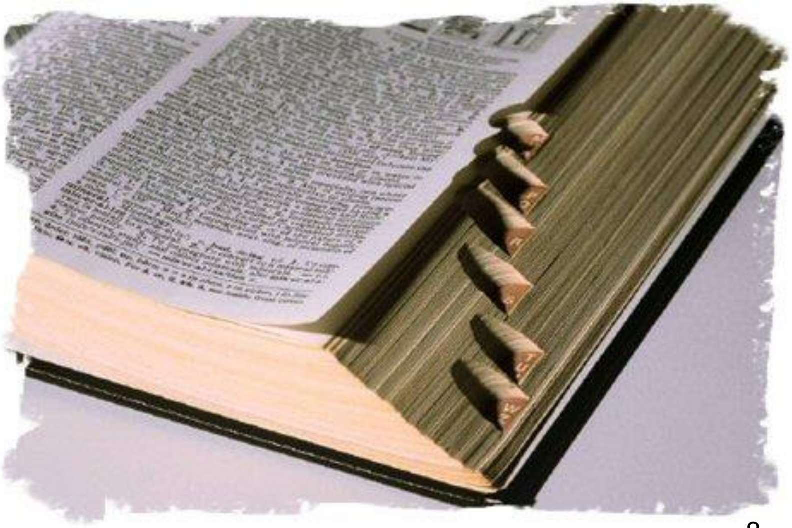
Par le Sieur DE HEDOVVILLE.

A PARIS,

Chez IEAN CVSSON, ruë S. Iacques, à l'Image de S. Iean Baptiste.

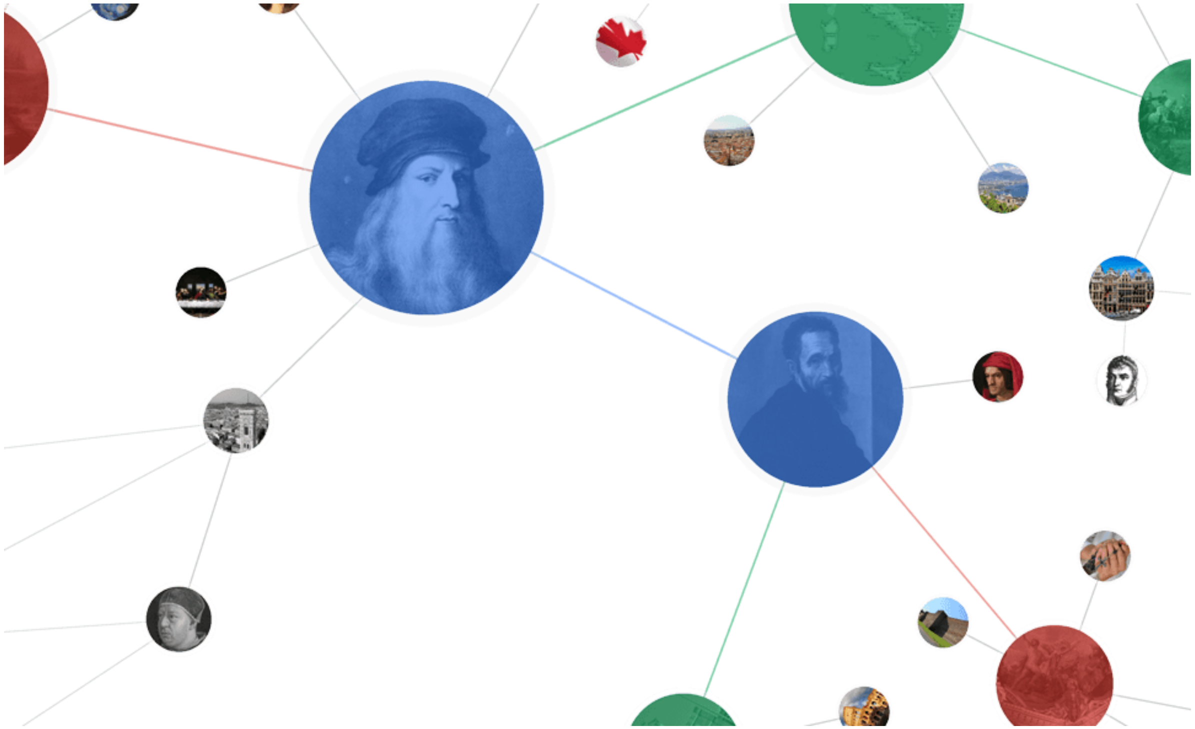
M. DC. LXV.

AVEC PRIVILEGE DV ROY.



Recent developments

- Google Knowledge Graph
  - Mining graph of entities (people, places, events, …), enhanced with semantic information



Some common features

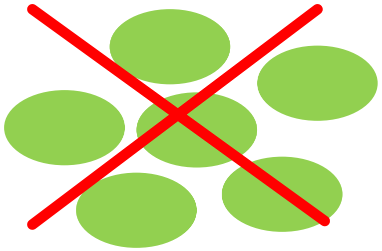
- Unconstrained
  - Any link can exist a-priori
- No rules or centralized design
  - Local decisions and incentives drive network formation
- Nobody has global information
  - Every actor typically knows its "neighborhood" only
- An element of chance
  - This suggests random models

Do these common features
give rise to common properties?

YES!

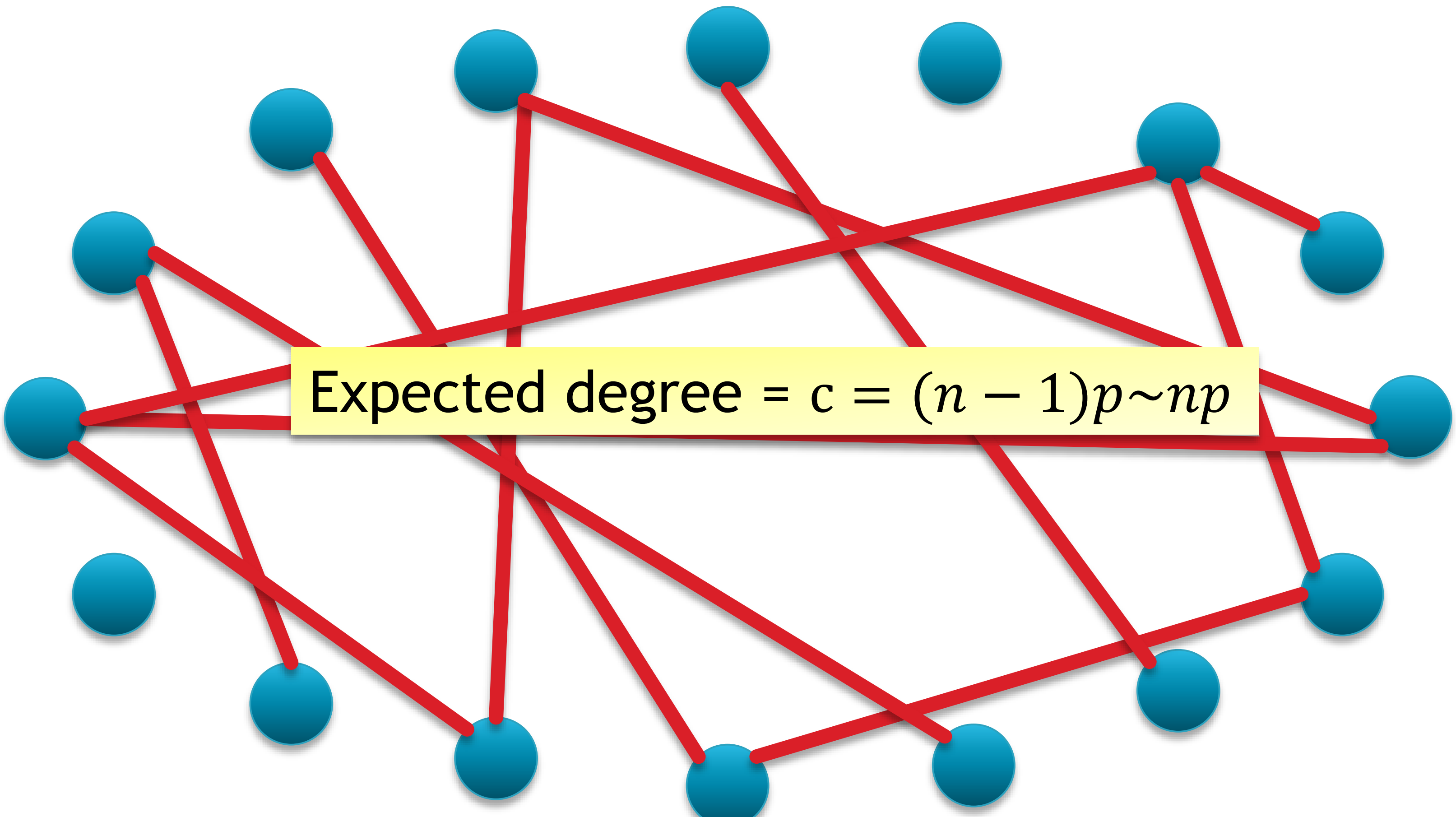
Property 1: giant component

- Definition:
  - Connected component that is
    - (a) much larger than other connected components
    - (b) significant fraction of whole network
- Q: why should there be a GC?
  - There could be several very large components
- Finding GC:
  - Finding component: start at vertex u, and run BFS/DFS to exhaustion
  - Find dominant component



Random graphs

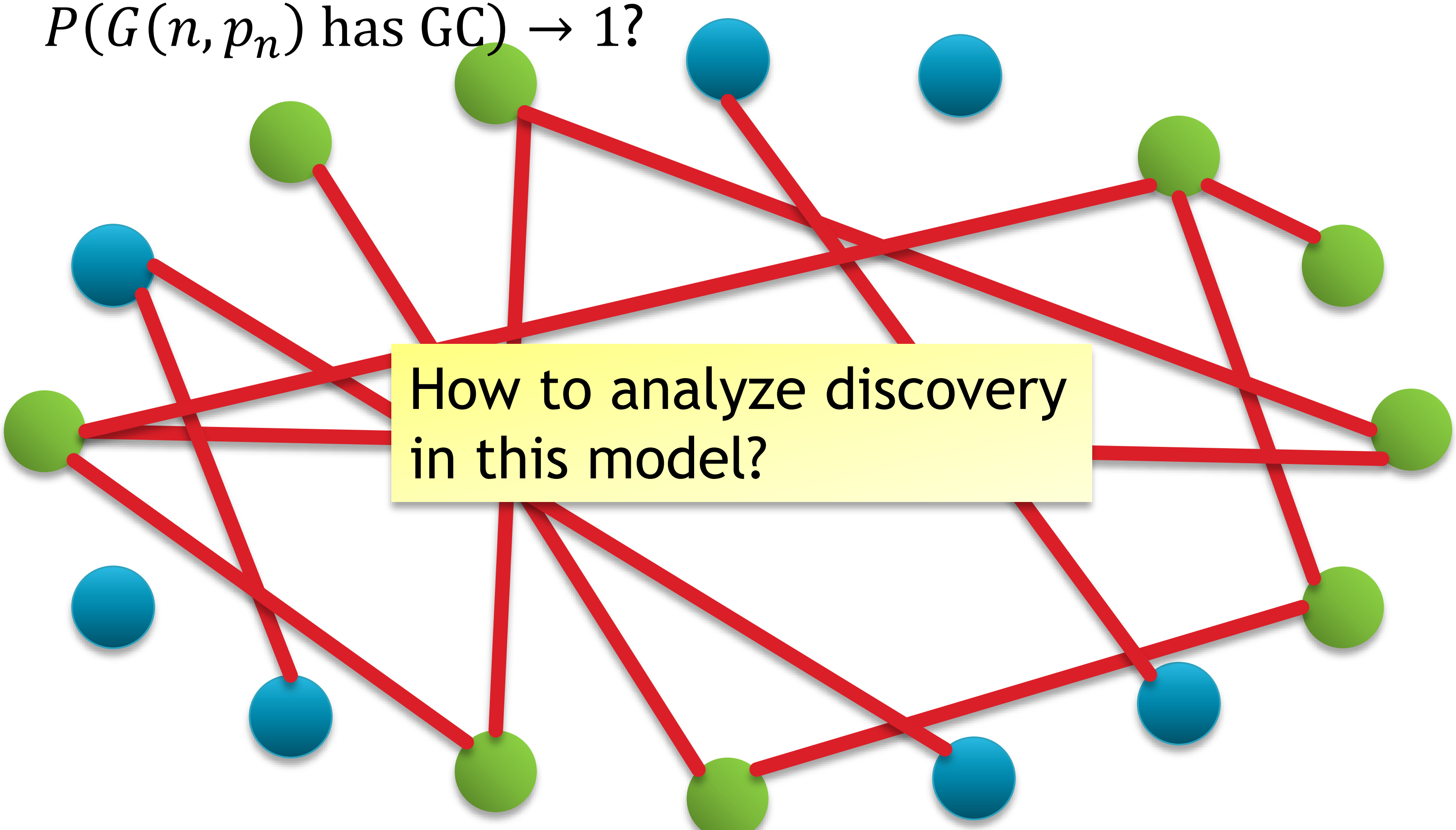
- $G(n,p)$ model:
  se genera el grafo G a partir de n y p
  - n vertices (usually very large)
  - Every edge (u,v) exists independently with prob. p



Expected degree = $c = (n-1)p \sim np$

Giant component in $G(n, p)$

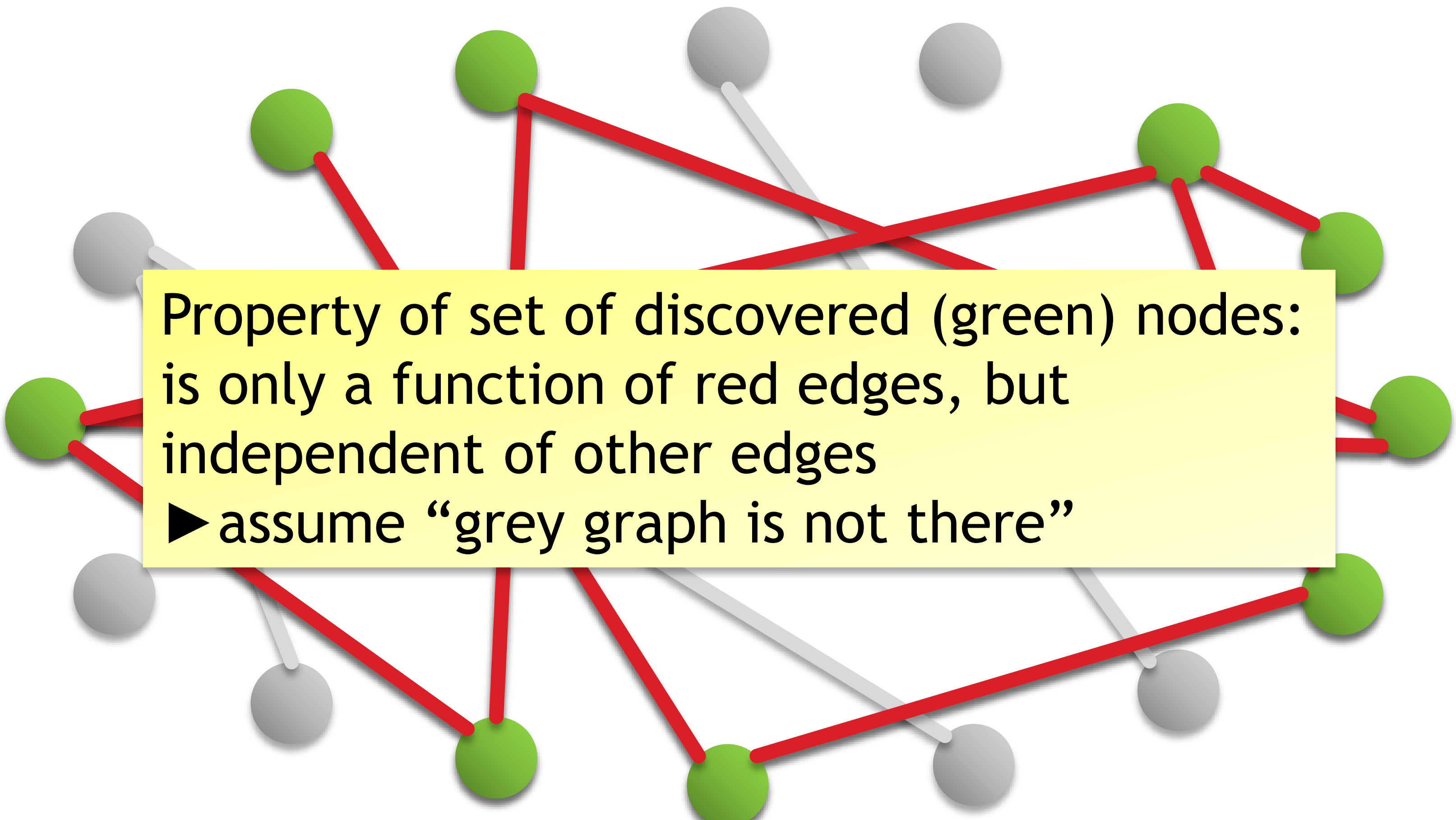
- Conditions for GC in $G(n, p)$?
- Precise question:
  - As $n \to \infty$, what functions p_n ensure that $P(G(n, p_n)\text{ has GC}) \to 1$?



How to analyze discovery in this model?

Giant component in $G(n, p)$

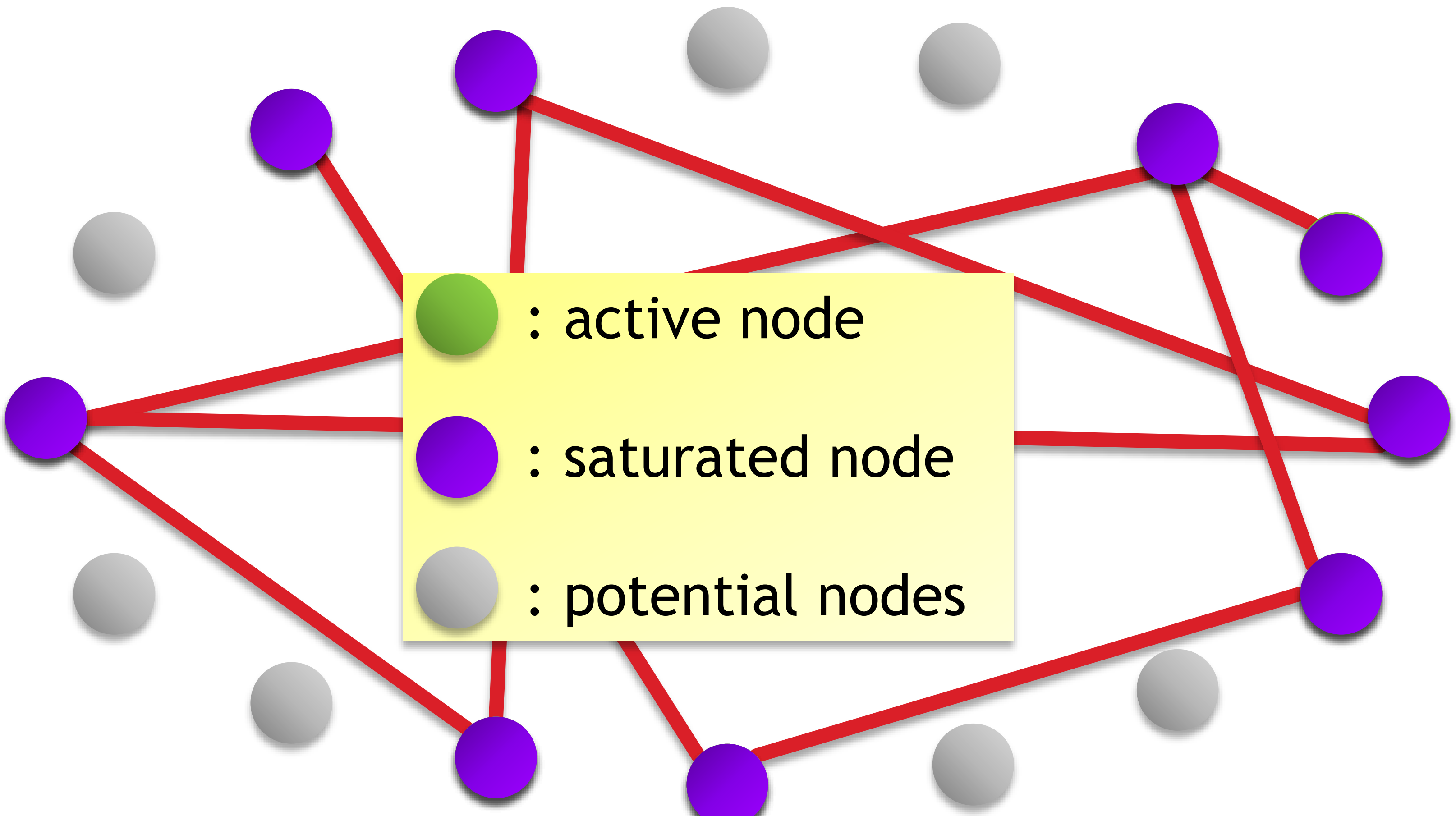
- Discovery process
  - Start at a node u
  - Find u's neighbors recursively until done (BFS/DFS)



Property of set of discovered (green) nodes:
is only a function of red edges, but
independent of other edges
►assume "grey graph is not there"

Giant component discovery in $G(n,p)$

- Using Principle of Deferred Decision (PDD)
  - Edges: flip coin only when needed



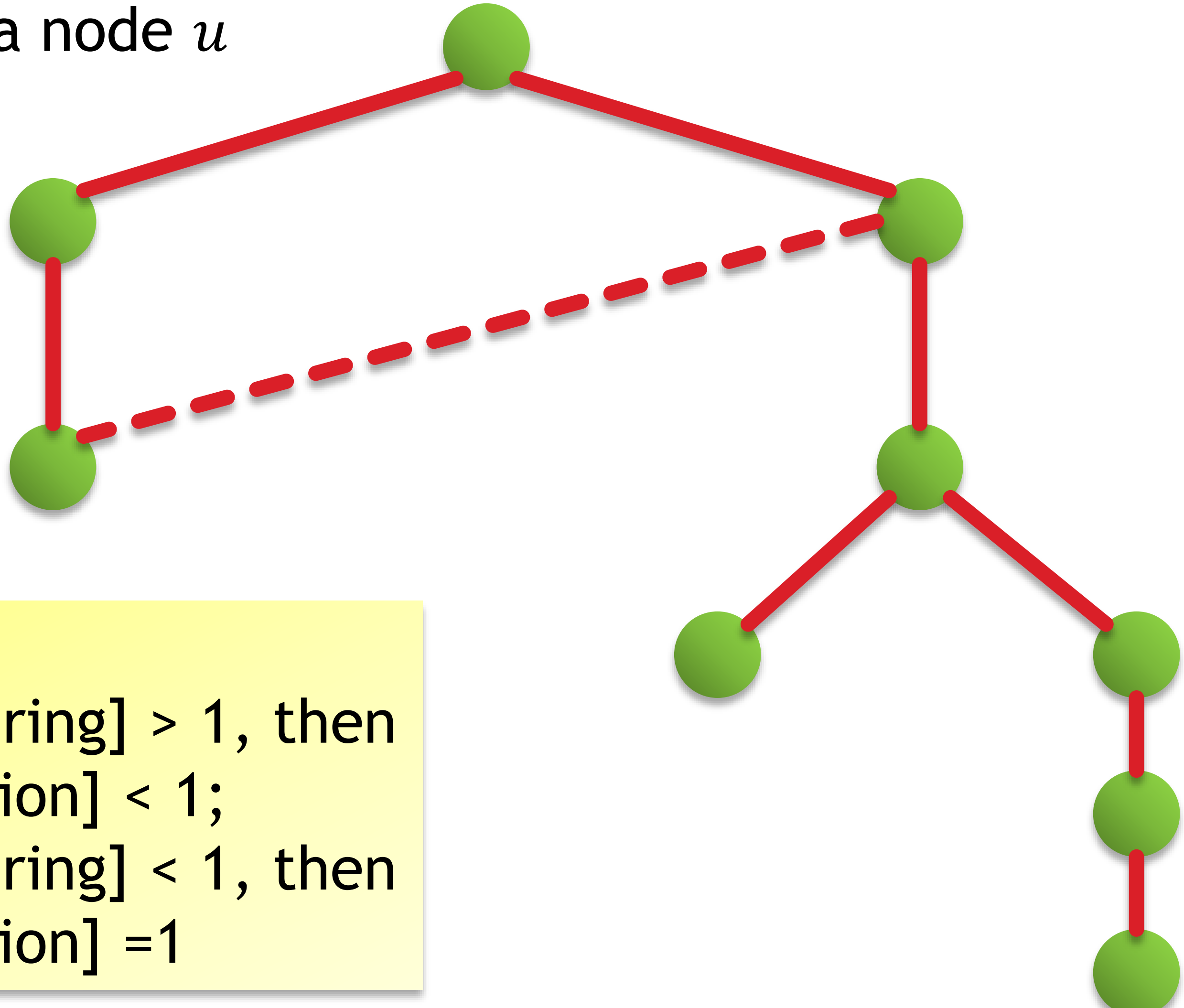
Giant component discovery

- kth step:
  - A_k: # active
  - k: # saturated (used)
  - $n - k - A_k$: # potential
- Number of new active nodes from old active node:
  - $X_k \sim \mathrm{Binom}(n - k - A_k, p)$
  - Independent
- Approximation:
  - While k and $A_k \ll n$,

$$\mathrm{Binom}(n - k - A_k, p) \cong \mathrm{Binom}(n, p)$$

Branching process: termination

- Discovery process:
  - Start at a node u



ModStoch:
If E[# offspring] > 1, then
P[termination] < 1;
If E[# offspring] < 1, then
P[termination] =1

Condition for GC in $G(n, p)$

- Set $p = \frac{c}{n}$
  - c: average degree
  - Number of offspring $\sim \mathrm{Binom}(n, \frac{n}{c}) \to \mathrm{Poisson}(c)$
- Theorem:
  - If $c > 1$, then $G(n, p)$ has a single component of size $\theta(n)$ asymptotically almost surely; all other components are small (of size $o(n)$)
  - If $c < 1$, then $G(n, p)$ has only small components

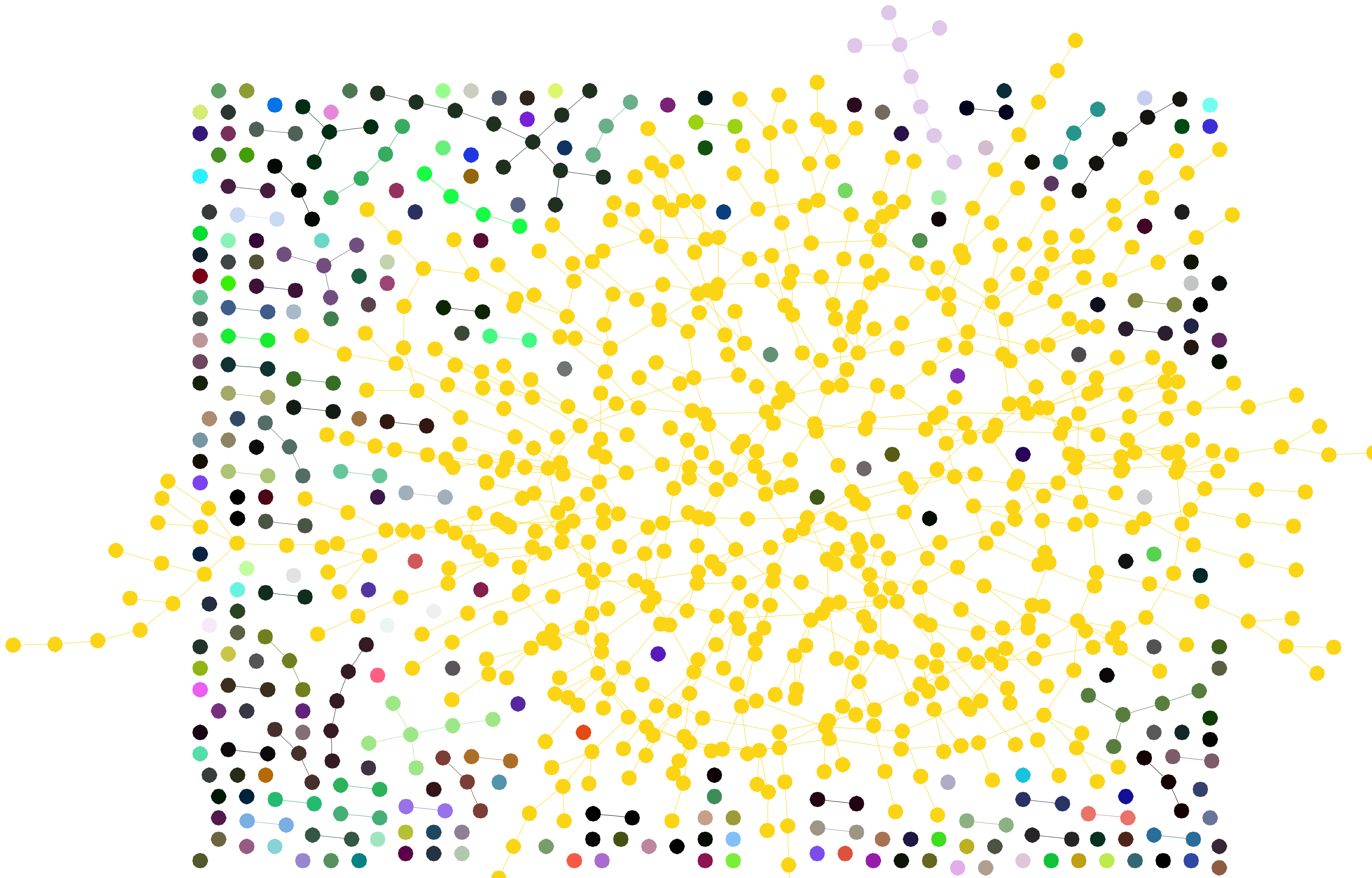
Condition for GC in $G(n, p)$

- Interpretation:
  - Giant component emerges naturally, even with completely random edge generation
  - No network-wide "coordination" needed
  - Sharp threshold – phase transition!
    - At avg degree $c = 1$

transicion escalonada
analogia con el agua y el hielo a los 0ºC

- Note:
  - More to prove: (i) that there's a **single** component; (ii) impact of ignoring $k + A_k$ negligible
  - Below threshold ($c < 1$): all small components are trees
  - Class "Models and Methods for Random Networks" for more details and properties of $G(n, p)$

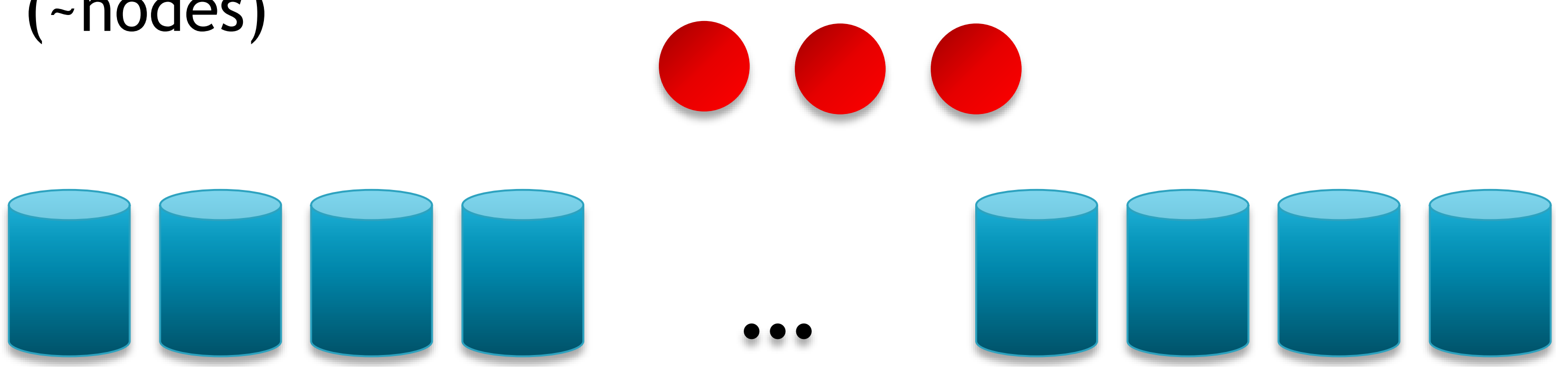
Example giant component



https://youtu.be/mpe44sTSoF8

$G(n,p)$ model: connectivity

- Another phase transition:
  - Consider threshold function $t(n) = \frac{\log n}{n}$
- Theorem:
  - If $p(n)/t(n) \to \infty$, then $G(n,p)$ is connected (a.a.s.)
  - If $p(n)/t(n) \to 0$, then $G(n,p)$ is not connected (a.a.s.)
  - Gap between these two → harder to analyze
- Intuition: Coupon collector problem
  - $n \log n$ balls (~edges) needed to have no empty bins (~nodes)



Coupon Collector problem

- Suppose $n-i$ empty bins, i bins have ≥ 1 ball
- X_i: # balls to go from i to $i+1$ filled bins
  - $X_i \sim Geom\left(\frac{n-i}{n}\right)$
- Z_n: # balls to fill all n bins
  - $Z_n = X_0 + X_1 + \cdots + X_{n-1}$
- $E[Z_n] = E[X_0] + E[X_1] + \cdots + E[X_{n-1}] =$

$$= \frac{n}{n} + \frac{n}{n-1} + \frac{n}{n-2} + \cdots + \frac{n}{1} =$$

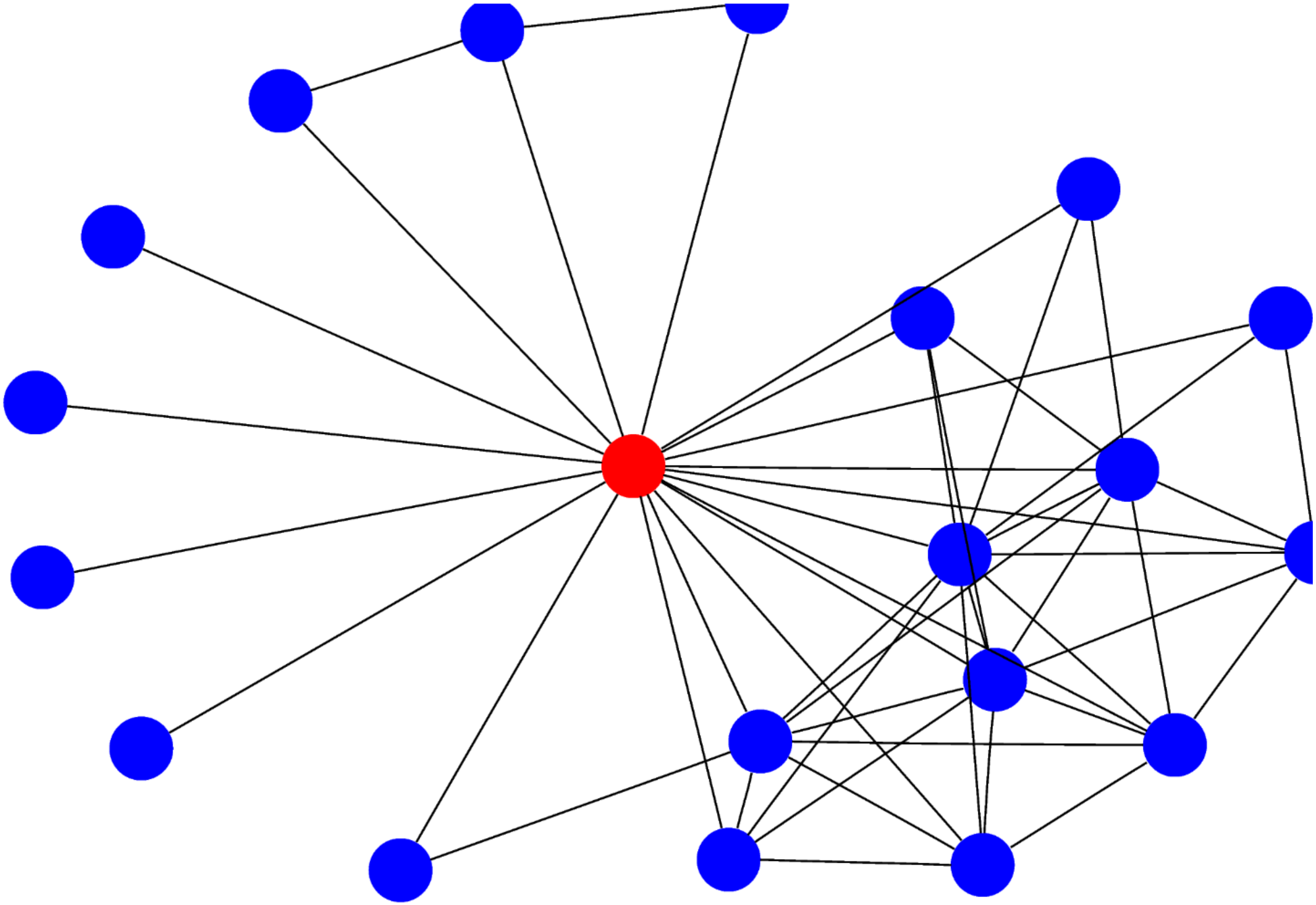
$$= n\left(1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n}\right) = \theta(n \log n)$$

- Back to $G(n,p)$: also need to show «single component» → beyond scope

$G(n, p)$ model: other phase transitions

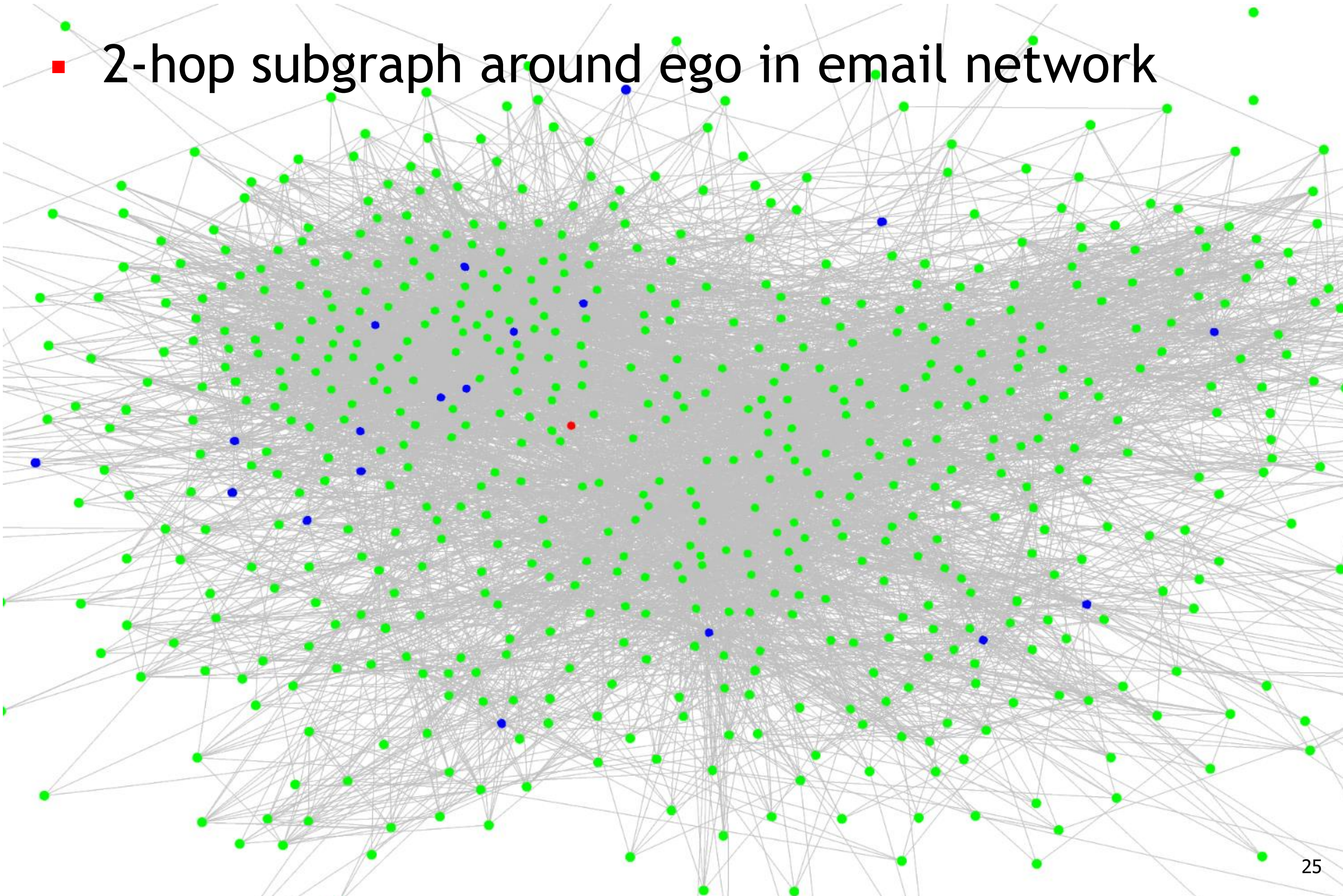
- First published 1959 by Erdös & Rényi
  - Focus on existence results
  - Very active field of research in probability
- Phase transitions:
  - Giant component
  - Connectivity
  - Existence of subgraphs
  - Chromatic number
  - Automorphism group
  - etc.

Property 2: clustering



Clustering

- 2-hop subgraph around ego in email network



Clustering metrics for single node

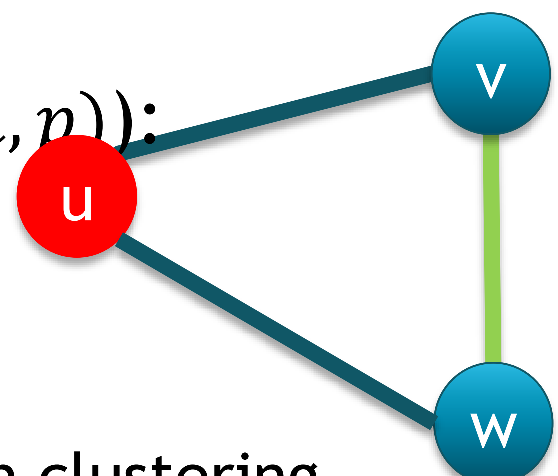
- Clustering = transitivity
  - Two nodes with common neighbor likely to be connected
- Clustering coefficient:
  - $c_u = \dfrac{|\{(v,w) \in E : (u,v) \in E, (u,w) \in E\}|}{\binom{d_u}{2}}$

    nº edges among friends (edges entre nodos que conectados a u)

    max nº edges among friends of u
  - = # links among friends / # possible links among friends
  - = empirical probability that (v,w) exists given (u,v) and (u,w) exist
- If links were entirely random ($G(n,p)$):
  - $E[c_u] = p \triangleq \dfrac{E[m]}{\binom{n}{2}} \cong \dfrac{2E[m]}{n^2}$

    expected coef

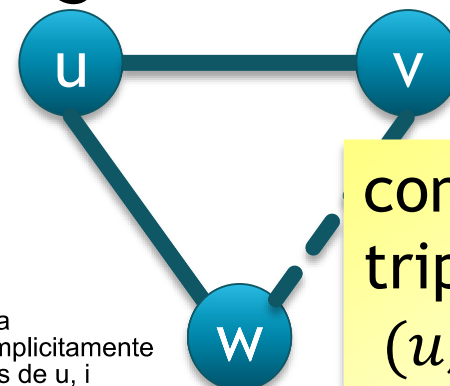
    where m = # edges in network
  - So $c_u \gg p$ means the network has high clustering



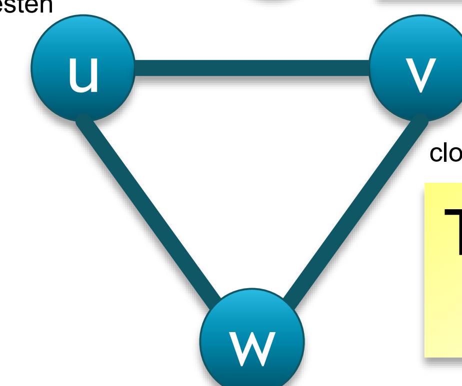
Clustering: two network-wide metrics

- Def 1: Average clustering coefficient
  - $c_G = \frac{1}{n} \sum_u c_u$
- Def 2: Weighted average clustering coefficient (also called "transitivity"):
  - $c_G = \dfrac{\sum_u \binom{d_u}{2} c_u}{\sum_u \binom{d_u}{2}} =$
  - $= \dfrac{\#\,\text{closed triples}}{\#\,\text{connected triples}}$
  - $= 3\,\dfrac{\#\,\text{triangles}}{\#\,\text{connected triples}}$

en connected triple se especifica el nodo que está en el centro. Implicitamente v y w estan contectados a través de u, i ndependientemente de que esten conectados tmb entre ellos



connected triple: $(u, \{v, w\})$



closed triple = triangle

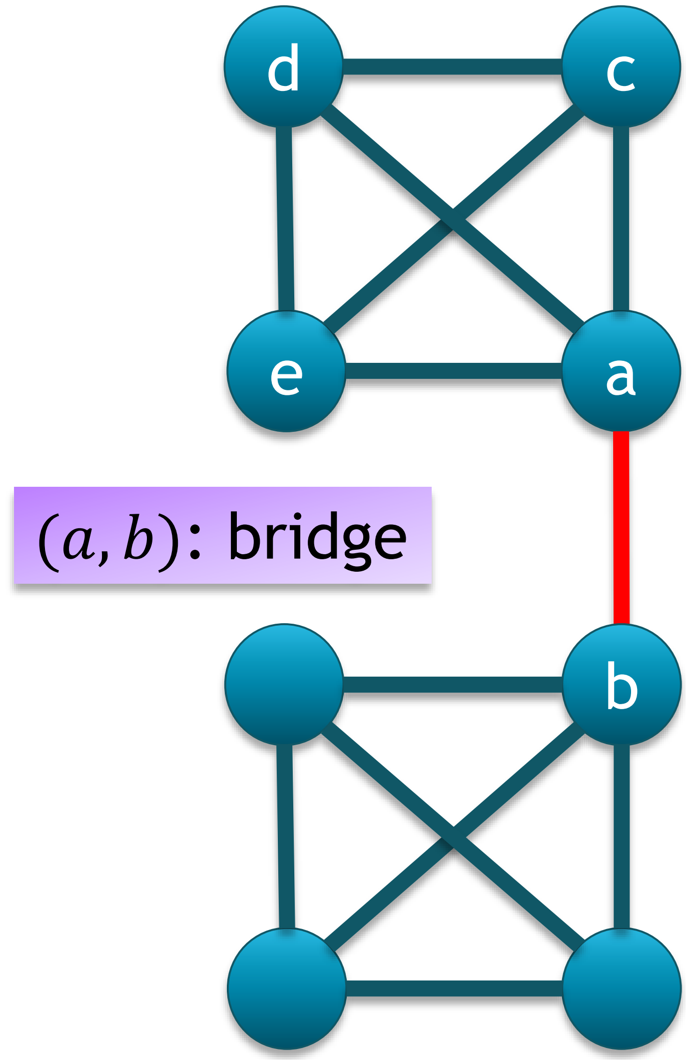
Triangle: $\{u, v, w\}$

Property 3: strong and weak ties

- Granovetter 1974: “The Strength of Weak Ties”
  - Observation from a survey:
    - Classify your social ties as either “strong” (close friends, family,…) or “weak” (acquaintances, professional colleagues,…)
    - If you found your last job through word-of-mouth, who told you?
  - Surprising result: job information came predominantly through weak ties!

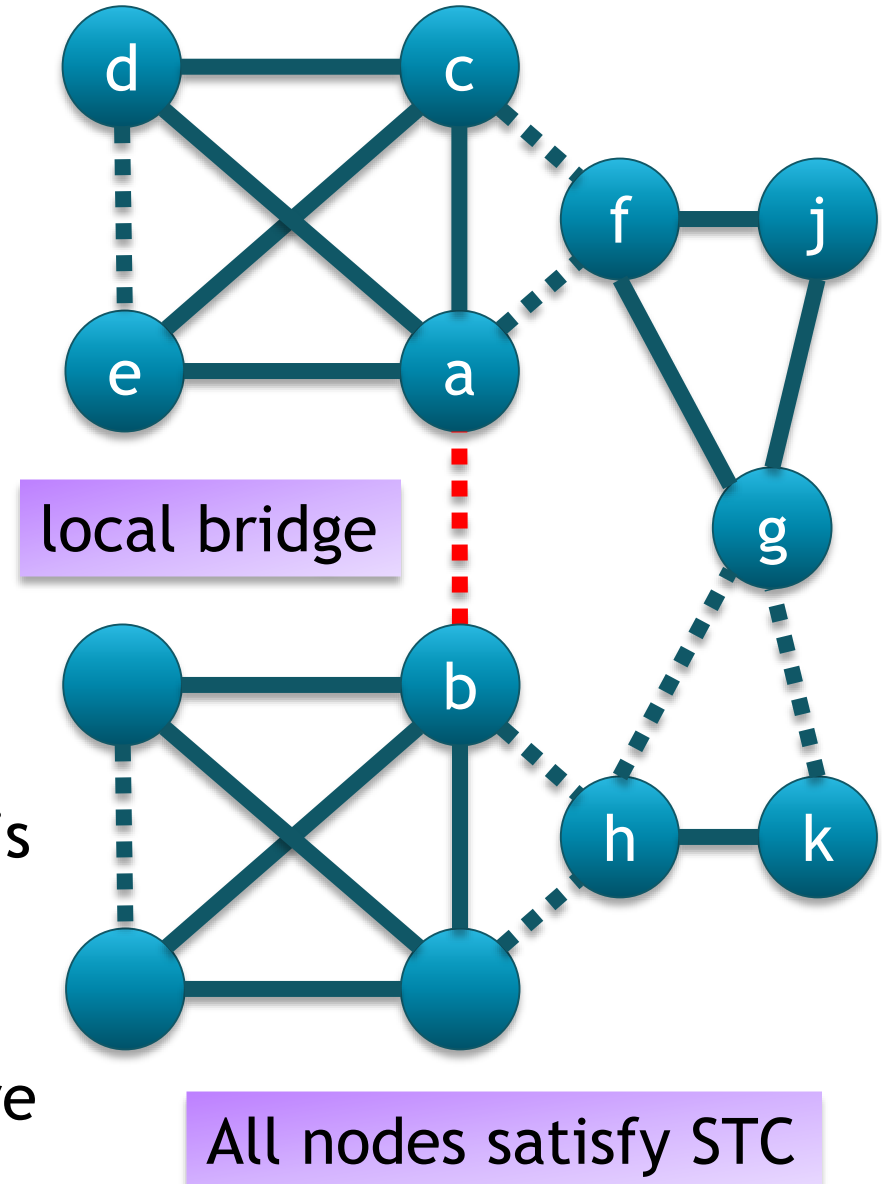
Bridges: "essential edges"

- Bridge (a, b):
  - Removing (a, b) disconnects
- Local bridge (a, b):
  - Removing (a, b) makes $d(a, b) > 2$
  - Equivalently: (a, b) have no common neighbors
- Informally:
  - "local bridge does not have short alternatives"



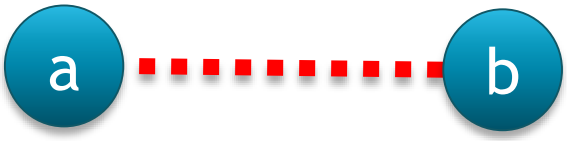
Strong and weak ties

- Edge property:
  - **Strong** tie: friendship, family, etc.
  - Weak tie: acquaintance, colleagues, etc.
- Strong Triadic Closure (STC) node property:
  - A node a violates STC if there are two **strong** edges (a, b) and (a, c), but there is no edge (b, c)
- Informally:
  - "two of a's **close** friends are likely to know each other"



STC ⇒ local bridges are weak ties

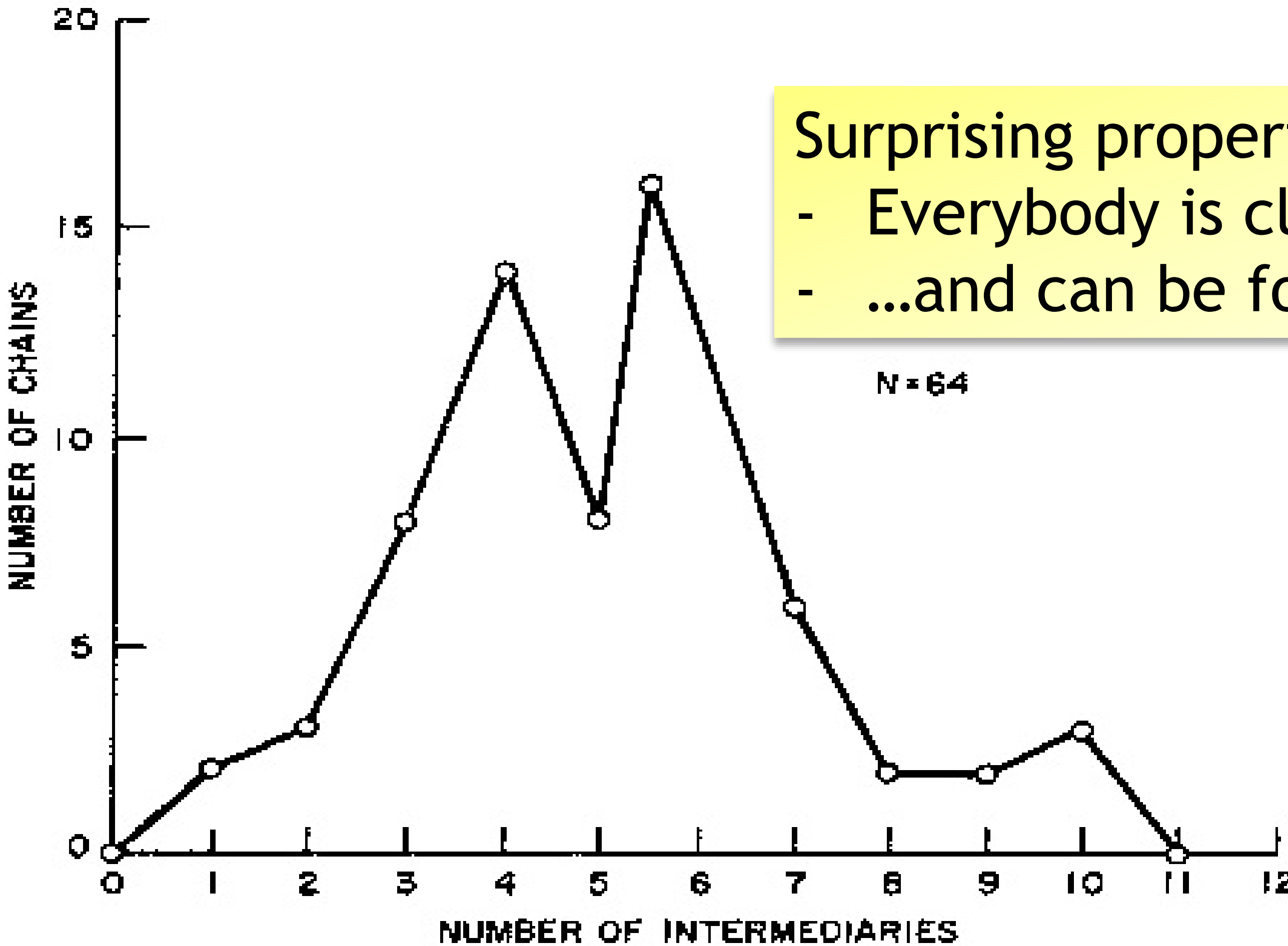
- Lemma:
  - If a node a satisfies STC (and has at least two strong ties), then any local bridge (a, b) is weak.
- Proof:
  - Assume node a satisfies STC, but (a, b) is strong and local bridge
  - By assumption, there is at least one other strong tie (a, c)
  - By STC, (b, c) must exist
  - But then a and b have common neighbor c, so (a, b) is not a local bridge
  - Contradiction



- Insight:
  - Social ties to other communities usually go through weak links

Property 4: short paths

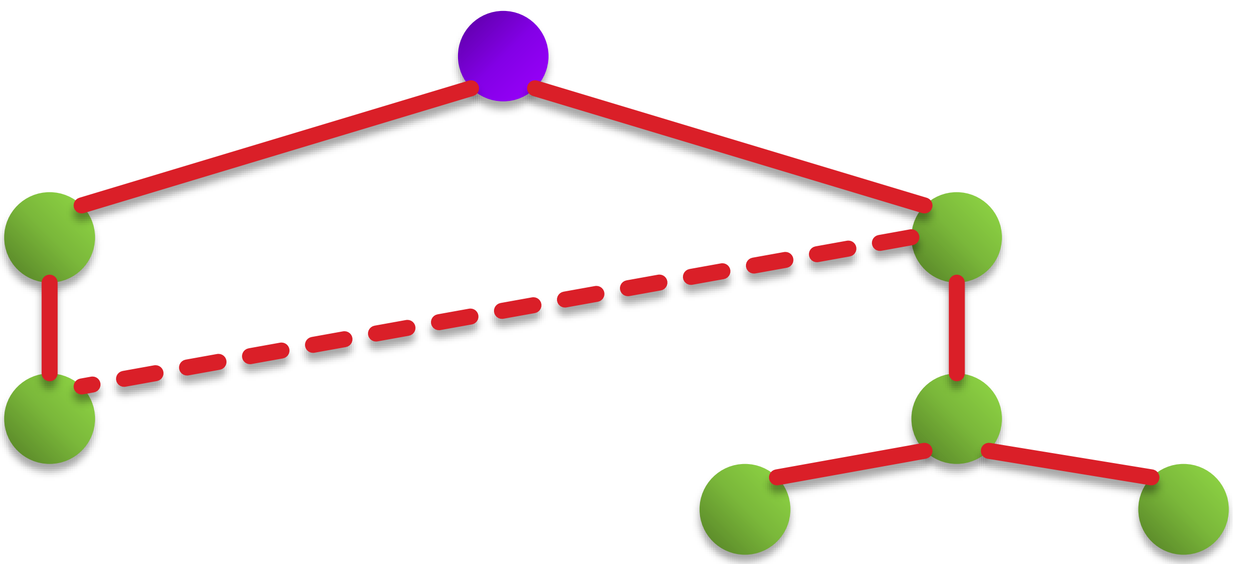
- Milgram 1969 experiment:
  - Letter passed through social links to find target



Surprising properties:
- Everybody is close
- ...and can be found!

Distances in random graphs

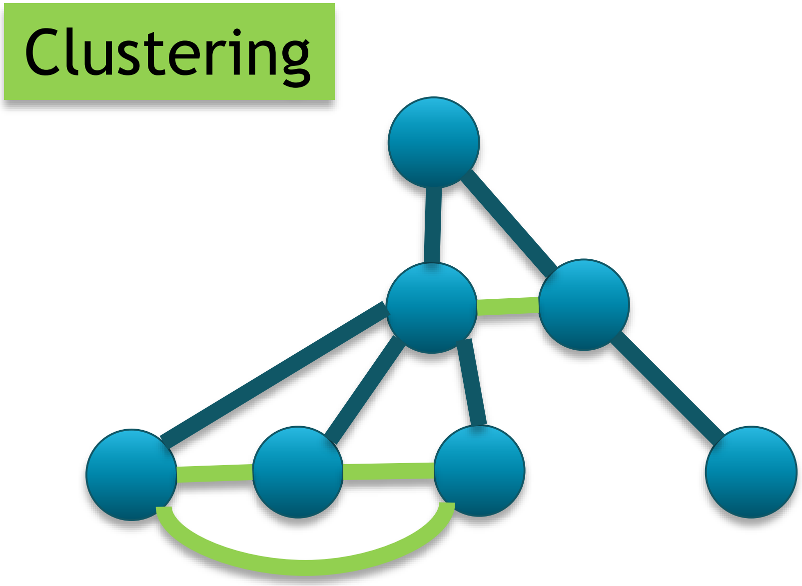
- Theorem (simplified):
  - $G(n, p)$ has diameter $\log(n) \;/\; \log(np)$
- Intuition:
  - The graph looks close to a tree from every node
  - Randomness creates very "efficient" graphs, i.e., edges used well to reach a large number of nodes
    - Few short cycles, incl. triangles

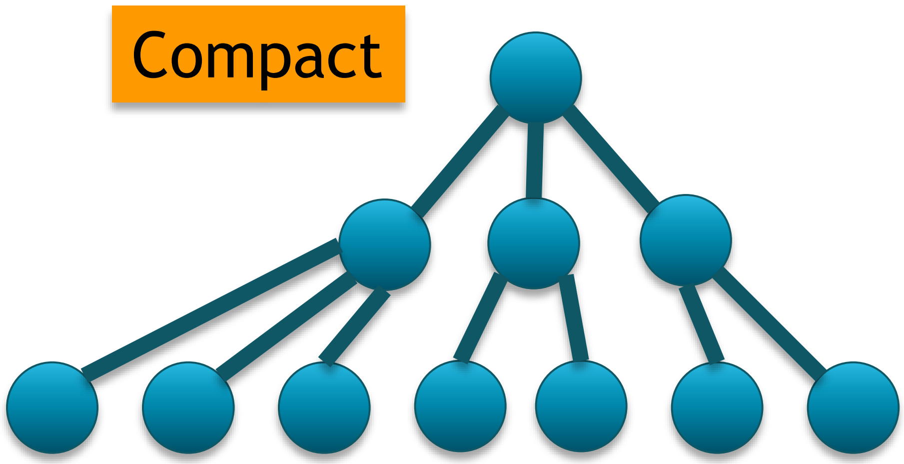


Recap: common network properties

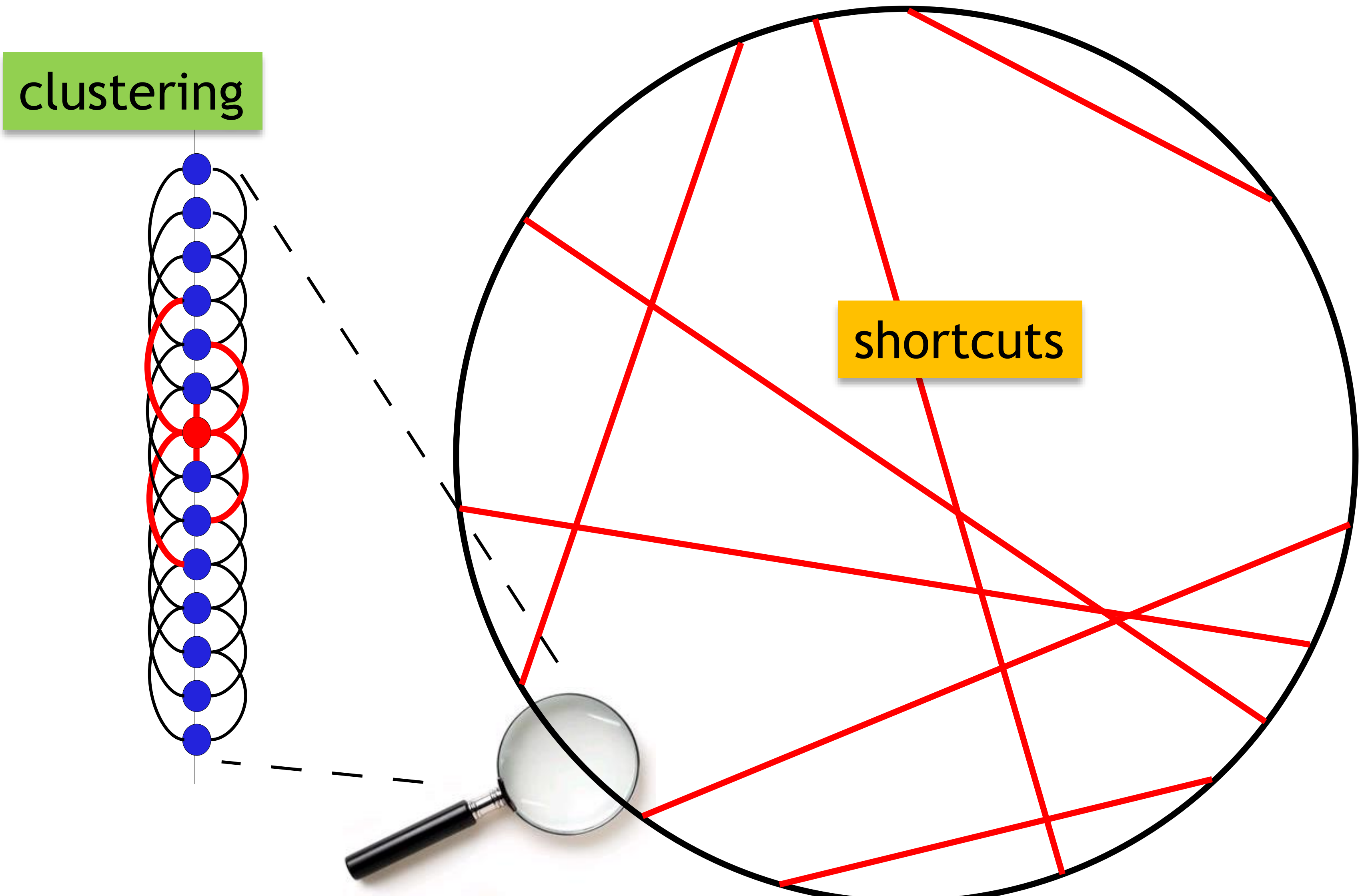
1	Giant component
2	Clustering
3	Strong and weak ties
4	Compact

(2) and (4) seem mutually exclusive: Paradox?



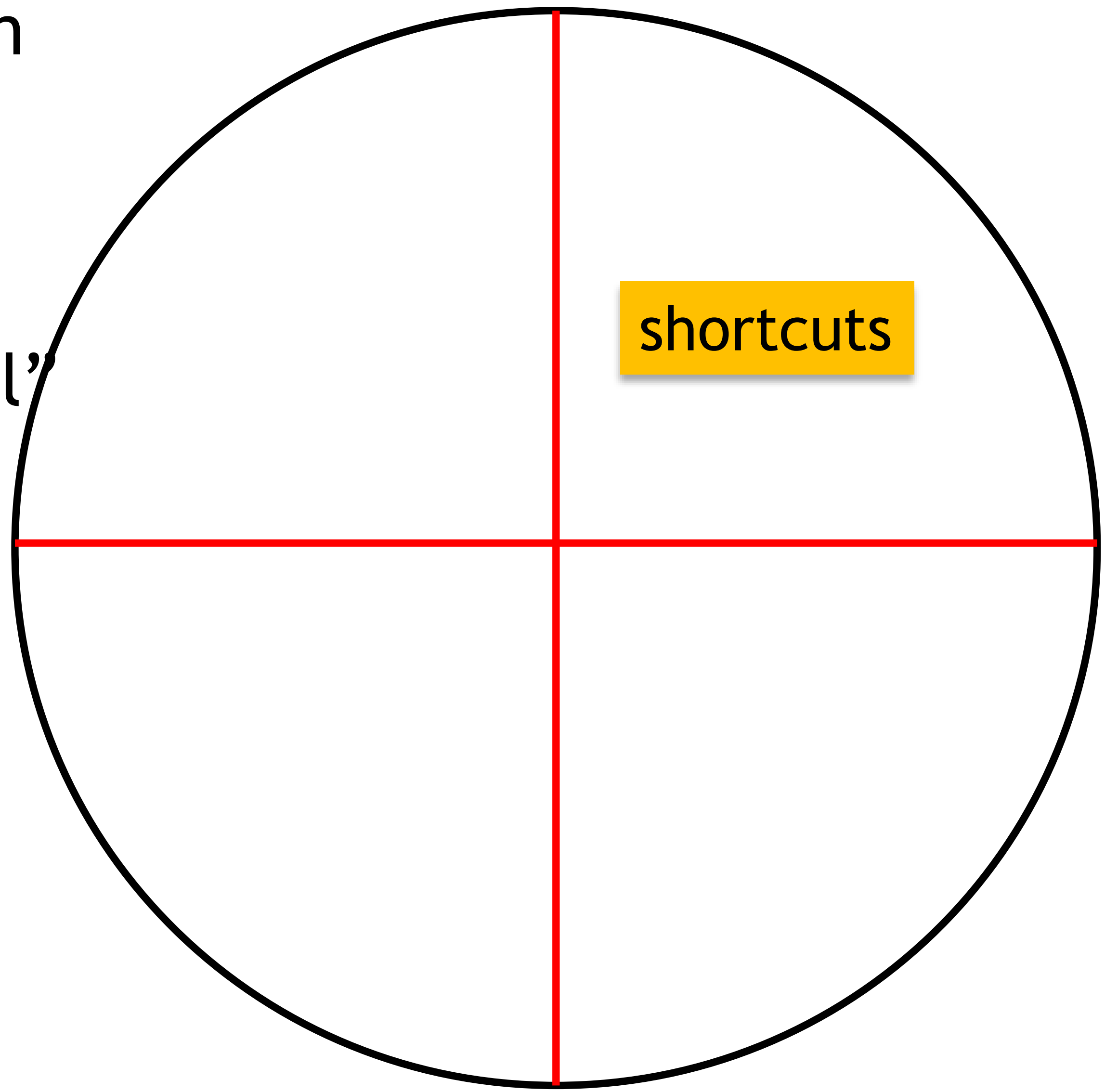


Small Worlds: Watts-Strogatz model



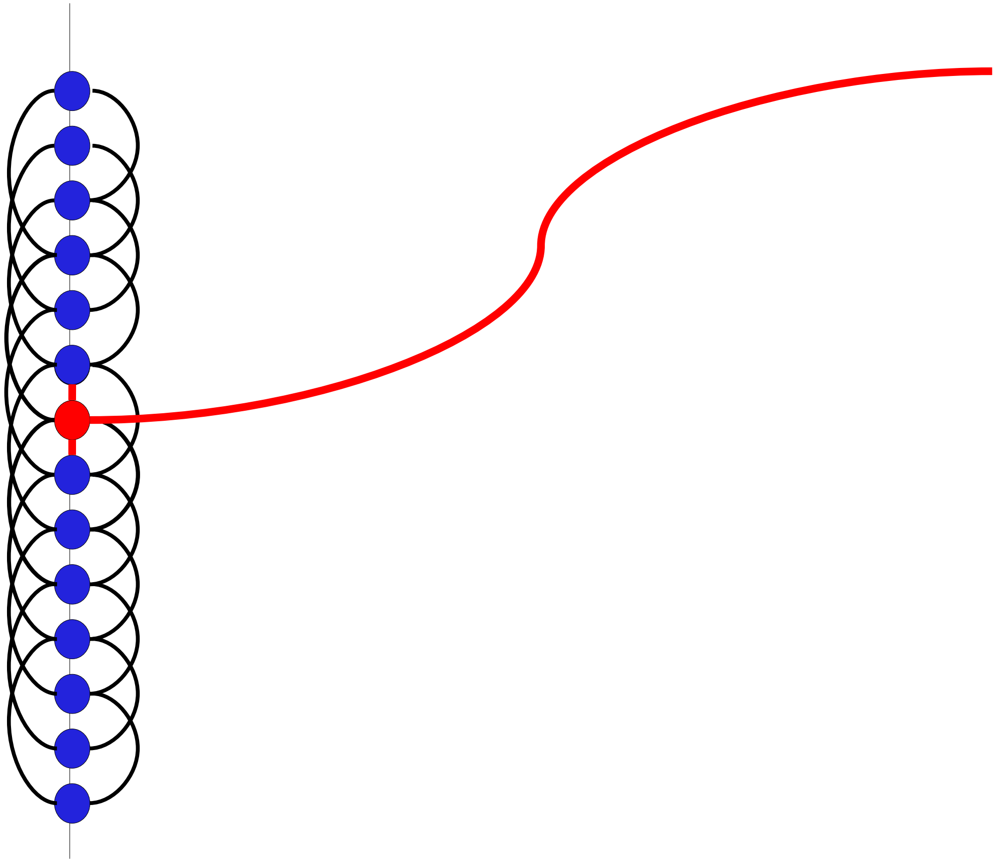
Evolution of avg distance with shortcuts

- Avg distance on circle:
  - cn
- Avg distance with one "ideal" shortcut:
  - $cn/2$
- With k shortcuts:
  - $O\left(\frac{n}{2^k}\right)$
- Distance drops quickly with k

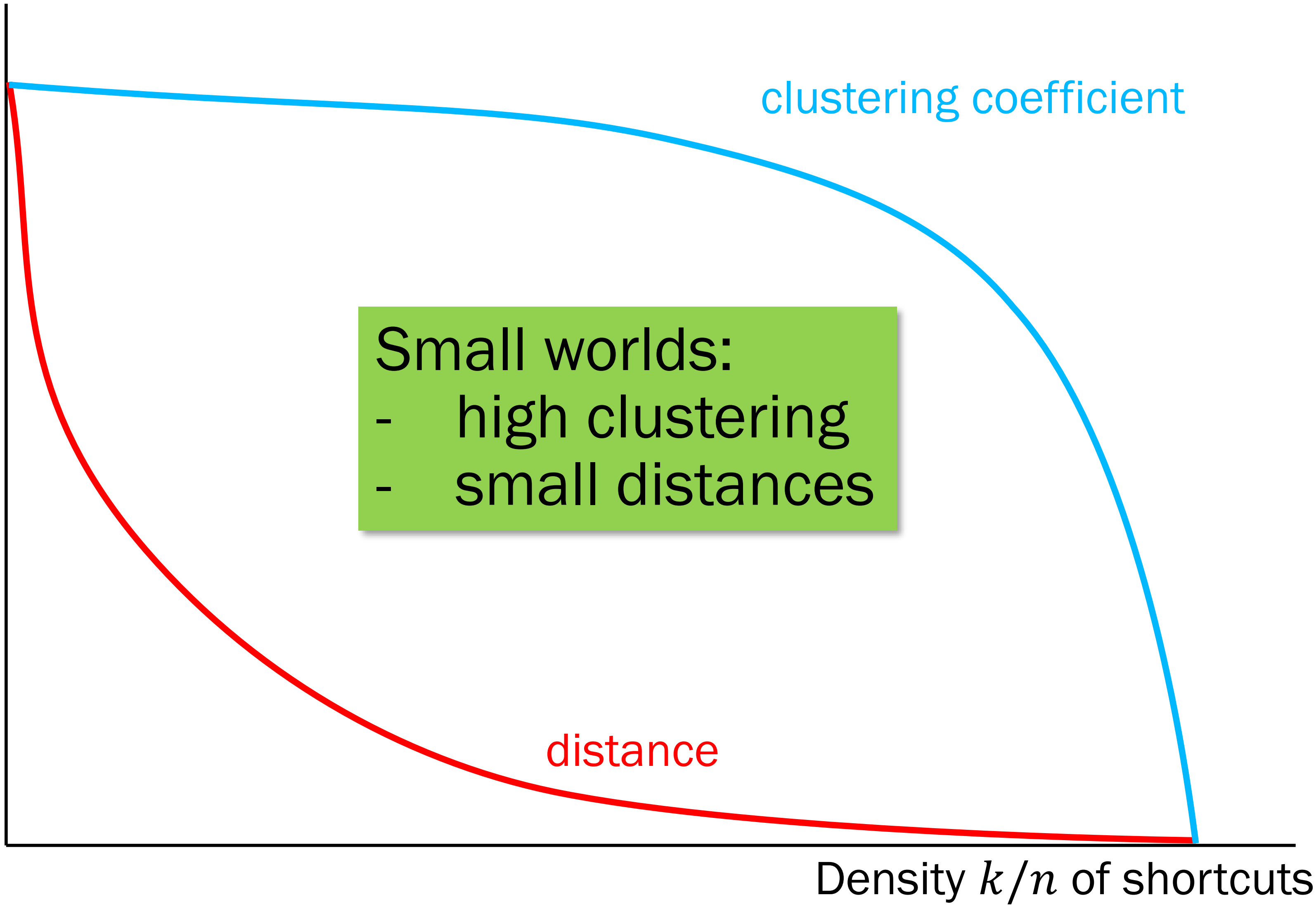


Evolution of clustering coeff with shortcuts

- As long as $k \ll n$, small impact on clustering coefficient
  - Number of potential triangles does not increase much



Clustering and diameter of WS



Summary & lessons

- Main properties of many types of “real world” and “self-organizing” networks:
  - Giant component: almost everything connected
  - Clustering: transitivity
  - Strong and weak ties: links connecting communities
  - Compact: everything is close
- Next week:
  - Network evolution and growth
- Lab objectives:
  - Explore real networks: degree distribution, giant component, small-world property
  - Using these properties, distinguish a road network from an internet graph
  - …and more related to next 3 lectures

References

- [D. Easley and J. Kleinberg: Networks, Crowds, and Markets (2010), chapter 3]
- [Bollobas, Random Graphs]
- [Newman, Networks]
- [Watts & Strogatz, Small Worlds]
- [Grossglauser & Thiran, Models and Methods for Random Networks (class notes), 2022]