Distribution-Free Models of Social and Information Networks

Tim Roughgarden (Stanford CS)

joint work with Jacob Fox (Stanford Math), Rishi Gupta (Stanford CS), C. Seshadhri (UC Santa Cruz), Fan Wei (Stanford Math), and Nicole Wein (Stanford CS)

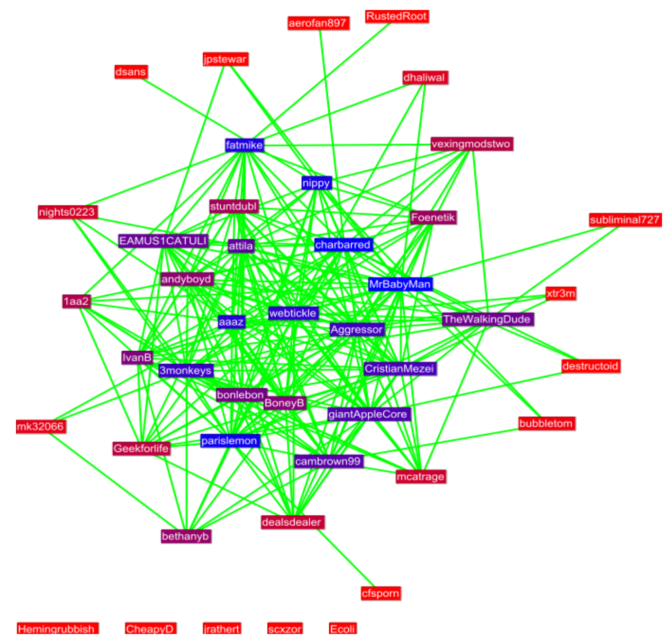
Social Networks Are Special

Consensus: Typical social and information networks have special structure.

- neither “worst-case” nor Erdös-Renyi

Examples:

- Facebook friendship graph
- Twitter follower graph
- LiveJournal
- Citation networks
- etc.

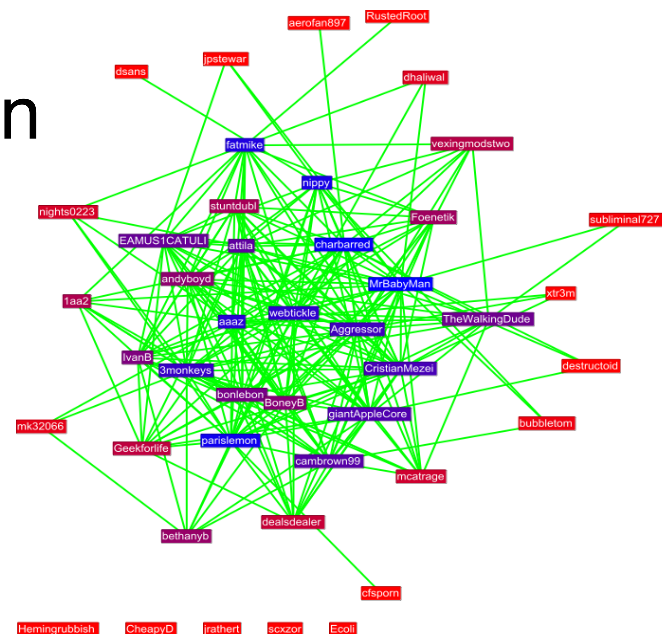


Social Networks Are Special

Partial consensus: Qualitative understanding of special structure.

Primary features:

- heavy-tailed degree distribution
- community structure
- triadic closure
- low-diameter/small-world



Social Networks Are Special

No consensus: Best way to articulate this structure.

Generative models: (cf., [Chakrabarti/Faloutsos 06])

- preferential attachment [Barabasi/Albert 99], …
- random graphs with given degrees [Chung-Lu 02], …
- copying models [Kumar et al. 99], [Kleinberg et al. 99], ...
- Kronecker graphs [Leskovec et al 10], …
- forest-fire model [Leskovec/Kleinberg/Faloutsos 07]
- BTER [Seshadhri/Kolda/Pinar 12]
- etc.

Motivating Questions

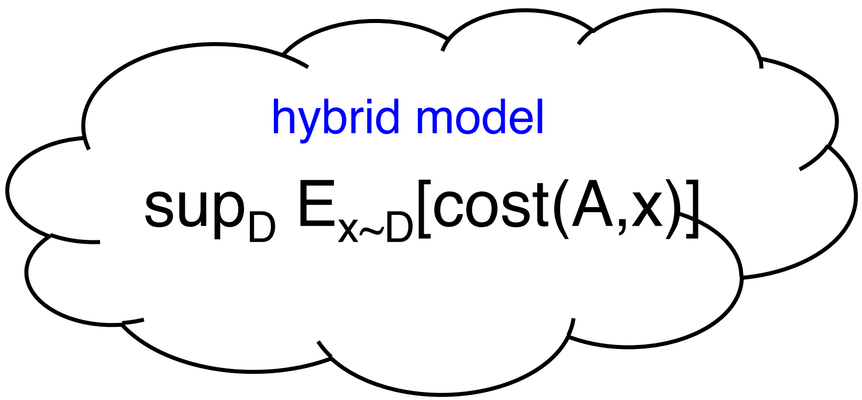
(i) what properties are shared by *all* good generative models of social and information networks? (present and presumably future)

(ii) do these minimal properties alone permit any interesting structural or algorithmic results?

Goal: results that hold simultaneously across all relevant generative models.

Inspiration: Hybrid Models

Motivation: for many problems there is a “sweet spot” between worst- and average-case analysis.

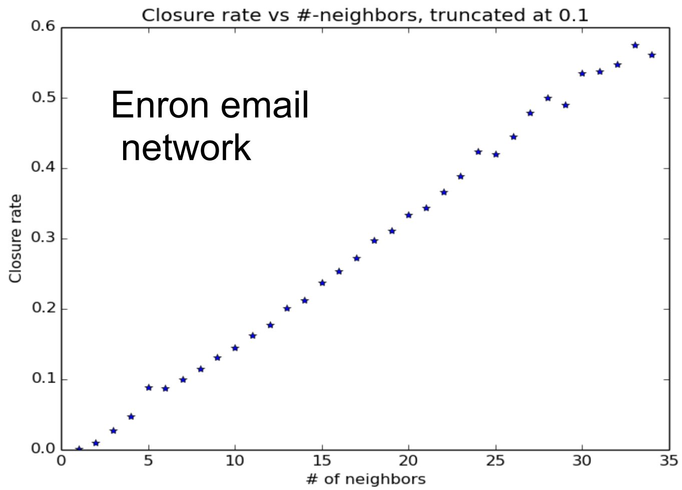


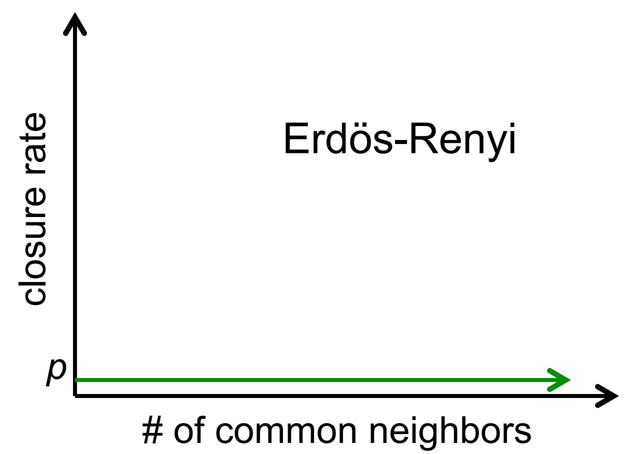
Example: smoothed analysis, semi-random models, prior-independent auctions, etc.

Key Property: Triadic Closure

Intuition: friends of friends likely to be friends themselves.

"It is argued that the degree of overlap between two individuals' friendship networks varies directly with the strength of their tie to one another." (M. S. Granovetter, "The Strength of Weak Ties," 1973)





On Power-Law Graphs

Fact: [Ferrante/Panduragan/Park 08] for typical NP-hard problems (e.g. clique) assuming only a power-law degree distribution doesn't help (much).

Recent developments: (for problems in P)

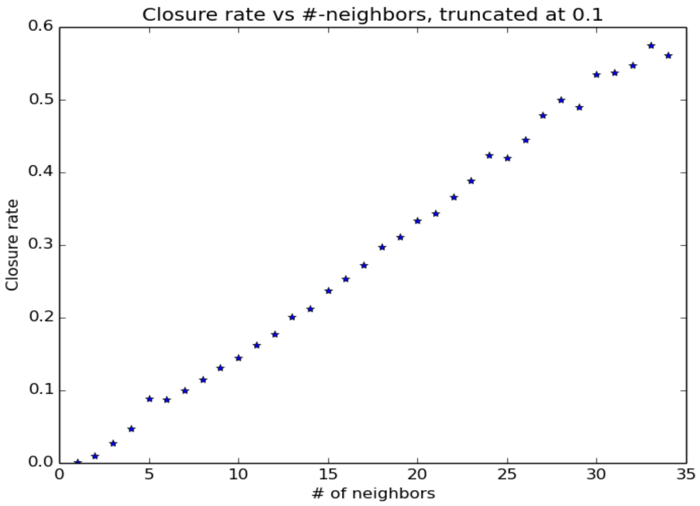
- [Brach/Cygan/Lacki/Sankowski 16]: faster algorithms for transitive closure, matching, etc.
- [Borassi/Crescenzi/Trevisan 16]: faster algorithms for diameter, radius, etc.

Part 1: Triangle-Dense Graphs

joint work with Rishi Gupta (Stanford CS)
and C. Seshadhri (UC Santa Cruz)
[appears in ITCS 2014, SICOMP 2016]

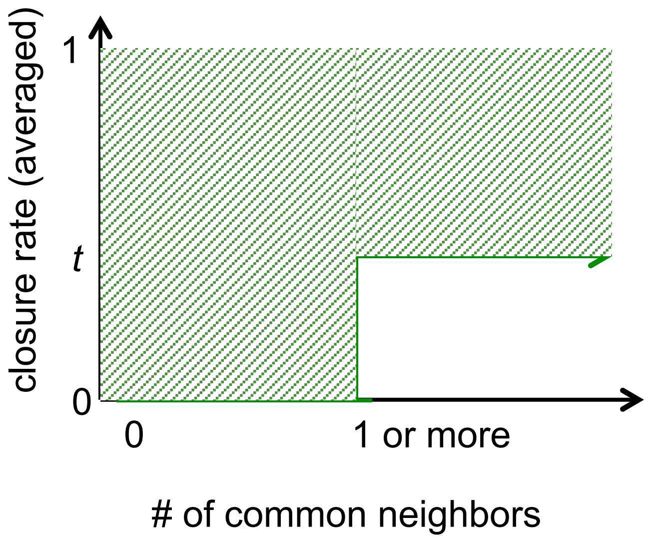
Triange-Dense Graphs (Cartoon)

The Data



average over pairs with at least one common neighbor

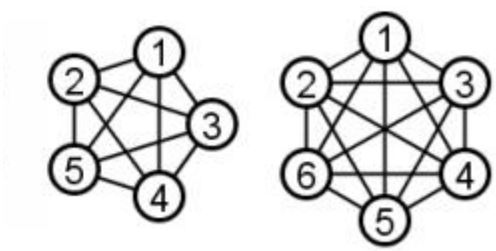
Our Mathematical Condition



Triangle-Dense Graphs

Definition: the *transitivity* of a graph := fraction of two-hops paths that are “filled in.”

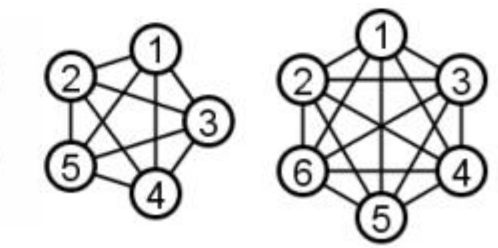
$$\frac{3\bullet(\#\text{ of } K_3\text{'s})}{\#\text{ of } K_{1,2}\text{'s}}$$



vs.

Triangle-Dense Graphs

Definition: the *transitivity* of a graph := fraction of two-hops paths that are "filled in."



vs.

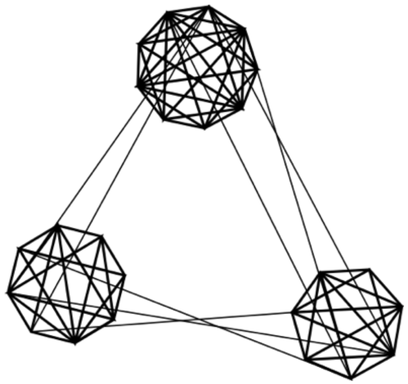
$$\frac{3\bullet(\#\text{ of }K_3\text{'s})}{\#\text{ of }K_{1,2}\text{'s}}$$

Definition: a family of graphs is *triangle dense* if the transitivity is at least a constant.

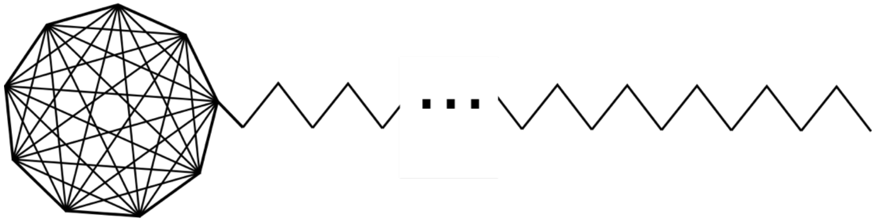
- Facebook transitivity ≈ .16 [Ugander et al. 11]
- edge density < 10^{-5}

What Do They Look Like?

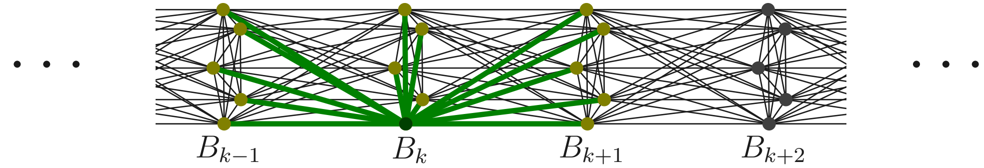
Idea: triangle-dense graphs = approximate unions of approximate cliques.



Issue: triangle-dense graphs can get weird.



- the “bracelet”:



Model of Social Networks?

Upshot: constant transitivity is a necessary but not sufficient condition to “look like a social network”.

Corollary: positive results for triangle-dense graph relevant for all good models of social networks.

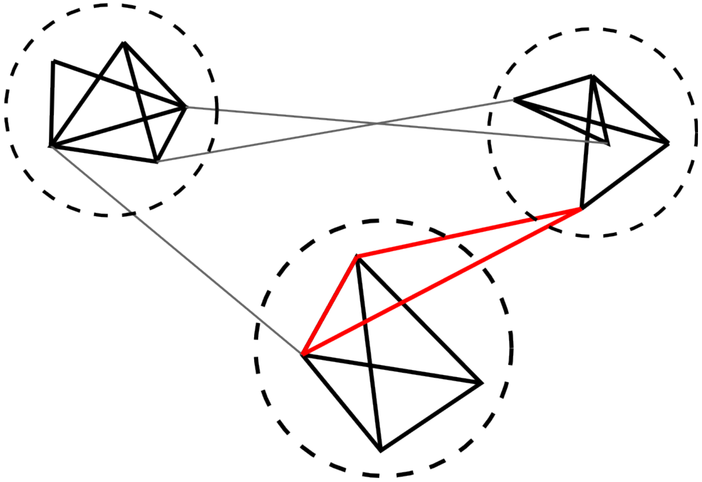
Worry: condition too weak for any positive results.

- cautionary tale: power-law degree distribution [Ferrante/Panduragan/Park 08]

A Decomposition Theorem

Theorem: a graph G with constant transitivity is approximately a union of cliques. Precisely, it contains a family of disjoint induced subgraphs s.t.:

- each has radius 2
- each is dense
  - in edges and triangles
- a constant fraction of G's triangles are preserved



(Parameters depend polynomially on the transitivity.)

Consequences

Conceptual: “community-like structures” are inevitable when transitivity is high.

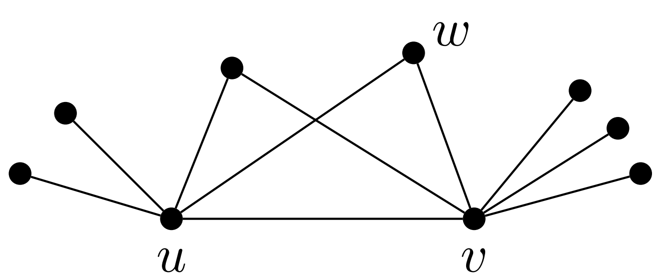
Algorithmic:

- proof gives a fast decomposition algorithm
- decomposition recovers ground truth clustering in “stable instance” model of [Balcan/Blum/Gupta 09]
- suggests a divide-and-conquer paradigm for triange-dense networks (e.g. for clique counting [Wang/Gupta/Seshadhri/Roughgarden 14])

Jaccard Similarity

Definition: *Jaccard similiarity* of an edge (u,v) =

$$|N(u) \cap N(v)| / (|N(u) \cup N(v)| - 2)$$



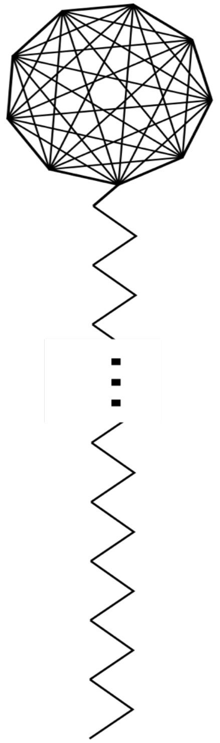
JS = 2/7

Note: equivalently, fraction of the wedges that (u,v) participates in that are triangles.

The Cleaner

Cleaner subroutine (ε): while there is an edge with Jaccard similarity < ε, delete it.

- ε = transitivity/4



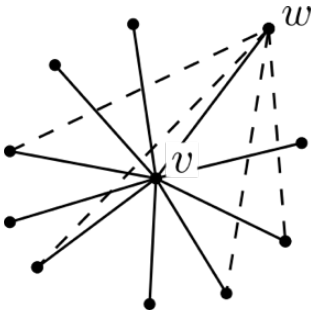
Lemma 1: the cleaner deletes at most a constant fraction of all triangles.

Reason: deletes many more wedges than triangles.

The Cleaner: Post-Conditions

Lemma 2: "locally approximately regular": the degrees of two adjacent vertices differ by at most a 1/ε factor.

Lemma 3: every 1-hop neighborhood $\{v\} \cup N(v)$ is edge- and triangle-dense.



Proofs: easy algebra.

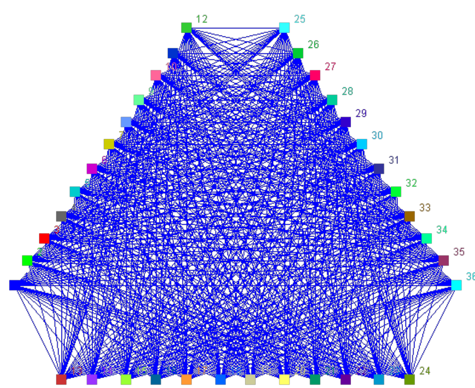
Candidate cluster: after cleaning, pick your favorite 1-hop neighborhood.

1-Hop Neighborhoods Fail

Candidate cluster: after cleaning, pick your favorite 1-hop neighborhood.

Problem: extracting a 1-hop neighborhood can destroy almost all triangles of the graph.

Bad example: complete tripartite graph.



The Extractor

Fix: greedily add vertices from the two-hop neighborhood until destroyed triangles can be "charged" to the saved triangles.

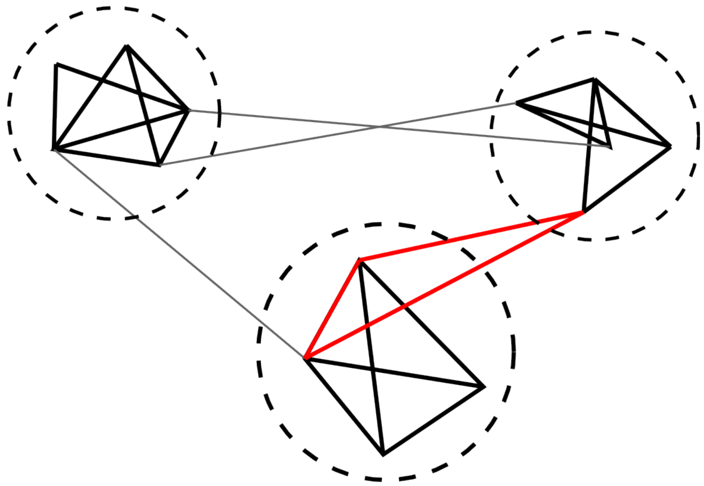
Non-trivial lemma: this can always be done.

Final algorithm: alternate the cleaning and extraction subroutines until graph is empty.

A Decomposition Theorem

Theorem: a graph G with constant transitivity is approximately a union of cliques. Precisely, it contains a family of disjoint induced subgraphs s.t.:

- each has radius 2
- each is dense
  - in edges and triangles
- a constant fraction of G's triangles are preserved



(Parameters depend polynomially on the transitivity.)

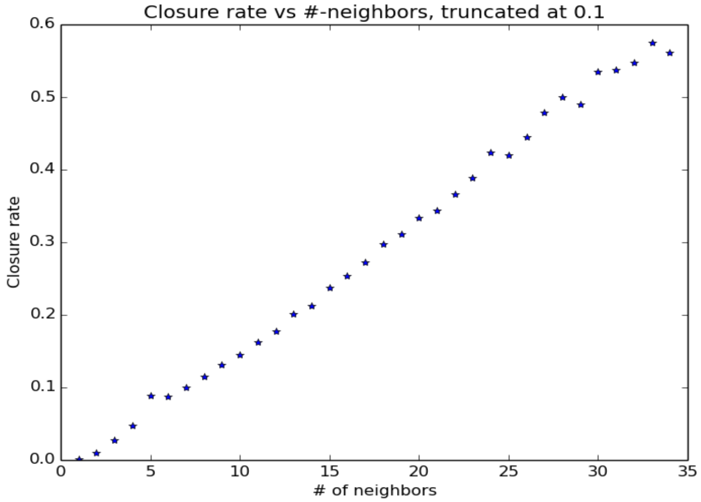
Part 2: c-Closed Graphs

joint work with Jacob Fox (Stanford Math), C. Seshadhri (UC Santa Cruz), Fan Wei (Stanford Math), and Nicole Wein (Stanford CS)

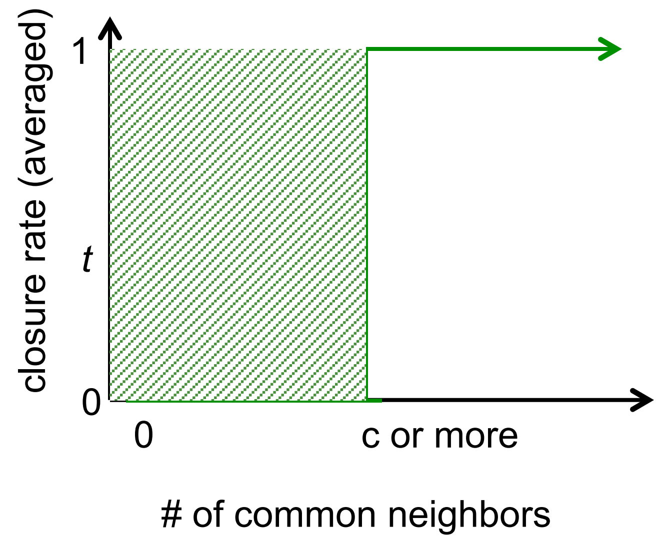
[ongoing work]

c-Closed Graphs (Cartoon)

The Data



Our Mathematical Condition



c-Closed Graphs

Granovetter (1973)

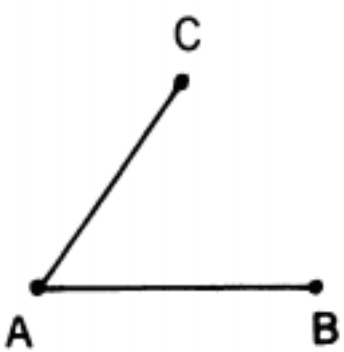
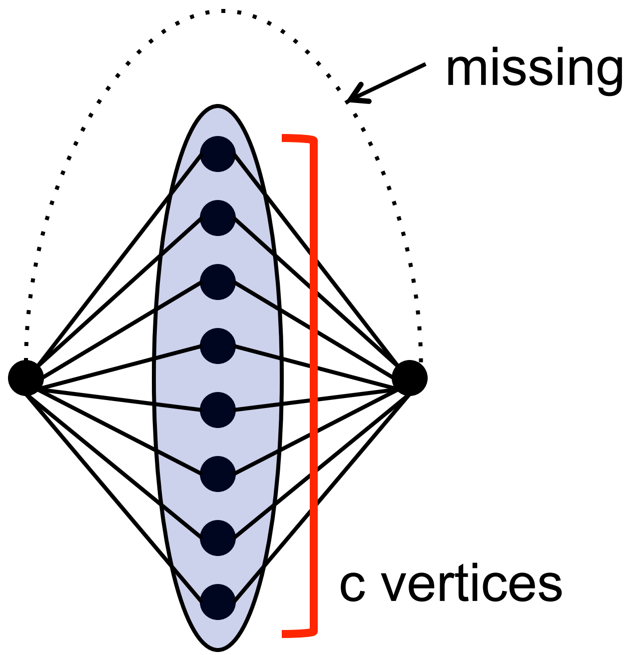


Fig. 1.—Forbidden triad

Our Forbidden Subgraphs



Definition: a graph G is *c-closed* if whenever u,v have ≥ c common neighbors, (u,v) is an edge.

c-Closed Graphs: Examples

Definition: a graph G is *c-closed* if whenever u,v have ≥ c common neighbors, (u,v) is an edge.

Example: disjoint union of cliques ⇔ 1-closed.

Example: girth ≥ 5 => 2-closed. (E.g., expanders.)

- plenty of hard problems stay hard on 2-closed graphs!

Question: but what about finding communities (e.g., the maximum clique)?

Enumerating Maximal Cliques

Fact: [Bron/Kerbosch 73] can enumerate maximal cliques in time polynomial in # of such cliques.

Bad example: An (n/3)-partite graph can have $3^{n/3}$ maximal cliques. (Tight by [Moon/Moser 65].)

- not c-closed for any $c \leq n-3$

Corollary: # of maximal cliques in a c-closed graph can scale exponentially with c.

Fixed-Parameter Tractability

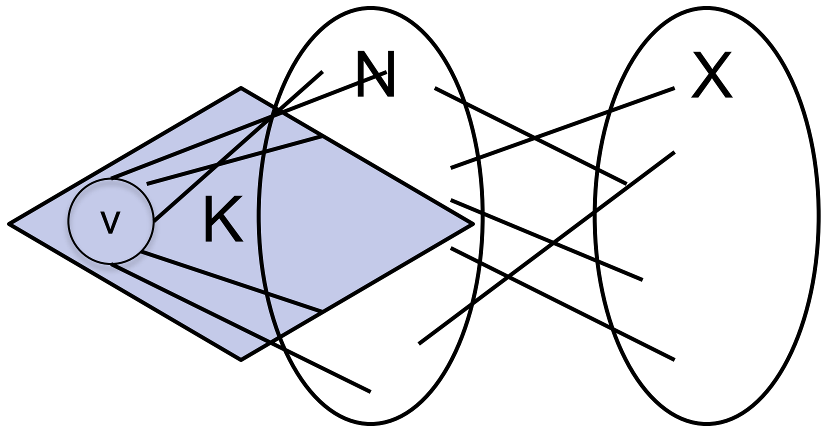
Theorem: the number of maximal cliques in a c-closed graph is at most $n^2\, 3^{c/3}$.

- dependence on n needs to be at least $n^{3/2}$

Corollary: for every fixed c, can solve max clique in polynomial time (even up to c = O(log n)).

Proof by Picture

Fix an arbitrary vertex v.

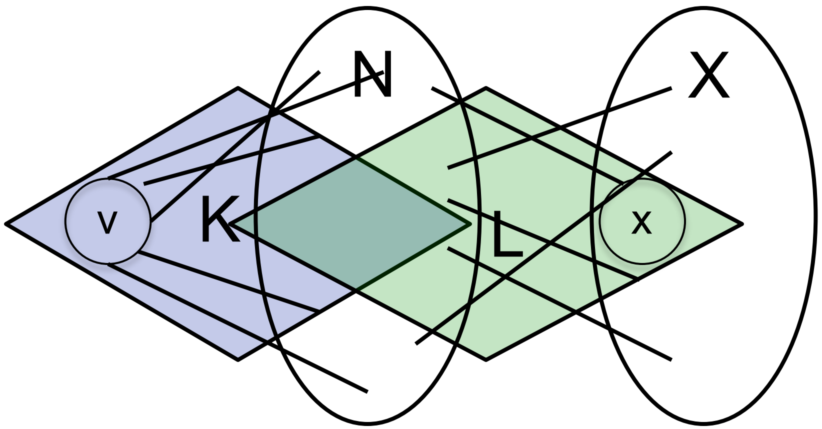


Case 1: cliques K where K\{v} maximal in G\{v}.

- induct on G\{v} (still c-closed)

Proof by Picture

Fix an arbitrary vertex v.



Case 2: cliques K where K\{v} *not* maximal in G\{v}.

- extend K\{v} to maximal clique L in G\{v}
- c-closed => each x in X has ≤ c-1 neighbors in N
- Moon-Moser => ≤ $3^{c/3}$ such L's per x in X

Some Open Questions

- quantitative improvements
  - decomposition dependence on triangle density
  - maximal clique dependence on graph size
- more algorithmic applications
  - possibly under stronger assumptions
- other “model-free” definitions of social networks
  - example: social networks have few induced squares [Ugander/Backstrom/Kleinberg 13]