Asymptotic Bounds for Clustering Problems in Random Networks

Eugene Lykhovyd and Sergiy Butenko

Department of Industrial and Systems Engineering
Texas A&M University
College Station, TX 77843
{lykhovyd,butenko}@tamu.edu

October 2, 2018, Gainesville, Florida

Outline

1. Introduction
2. Clique relaxations
3. Asymptotic bounds in random networks
4. Conclusion

Outline

1. Introduction
2. Clique relaxations
3. Asymptotic bounds in random networks
4. Conclusion

Network-based analysis of big data

Big data arising in various complex systems can be conveniently modeled using networks/graphs:

- components of the complex system – vertices
- pairwise interactions between different components – edges

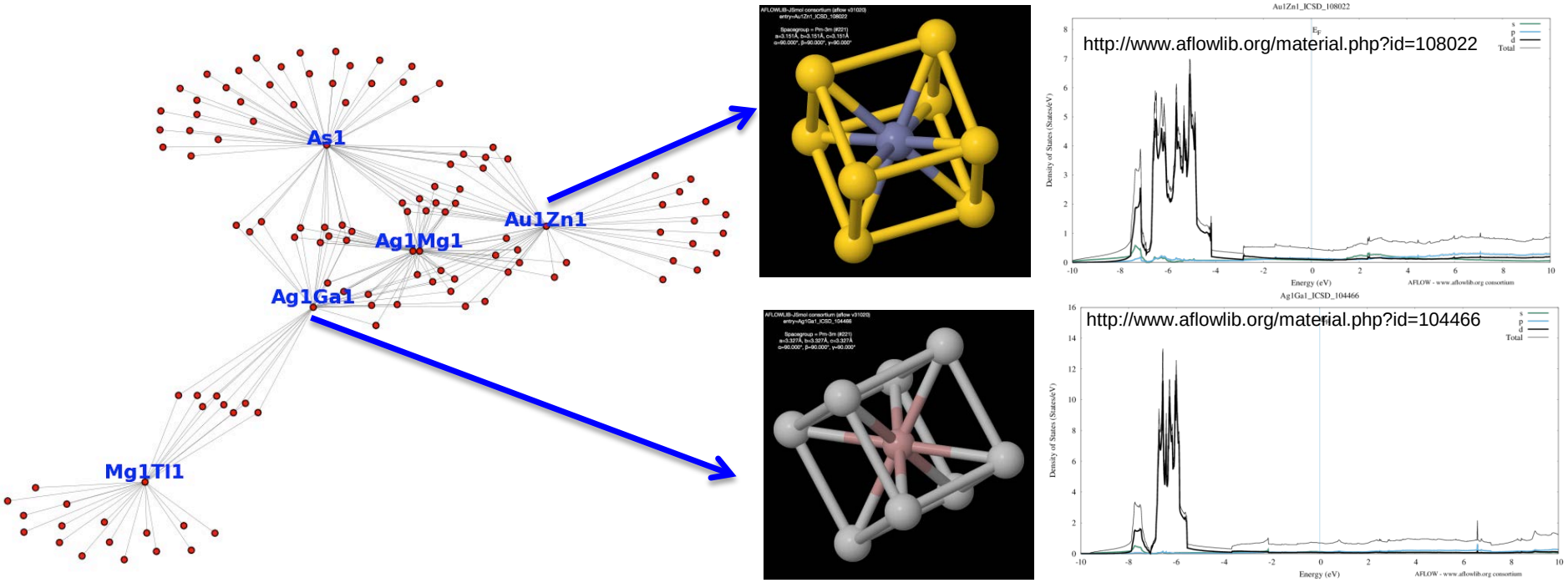
Network-based analysis allows to capture some global structural properties of the system and predict overall trends in its dynamics.

Clusters in networks

In many applications one is interested in detecting/designing cohesive clusters

- Communication/sensor networks
- Transportation/supply networks
- Power grid
- Cybersecurity
- Biological networks
- Social networks
- Financial networks

Networks of materials



A. Veremyev, L. Liyanage, V. Boginski, M. Fornari, M. Buongiorno Nardelli, S. Curtarolo, and S. Butenko. Networks of materials. Working paper.

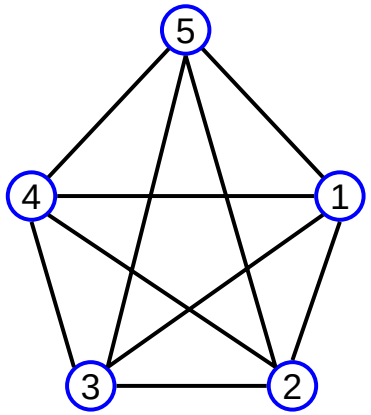
Social networks

A social network is described by $G = (V, E)$ where V is the set of "actors" and E is the set of "ties".

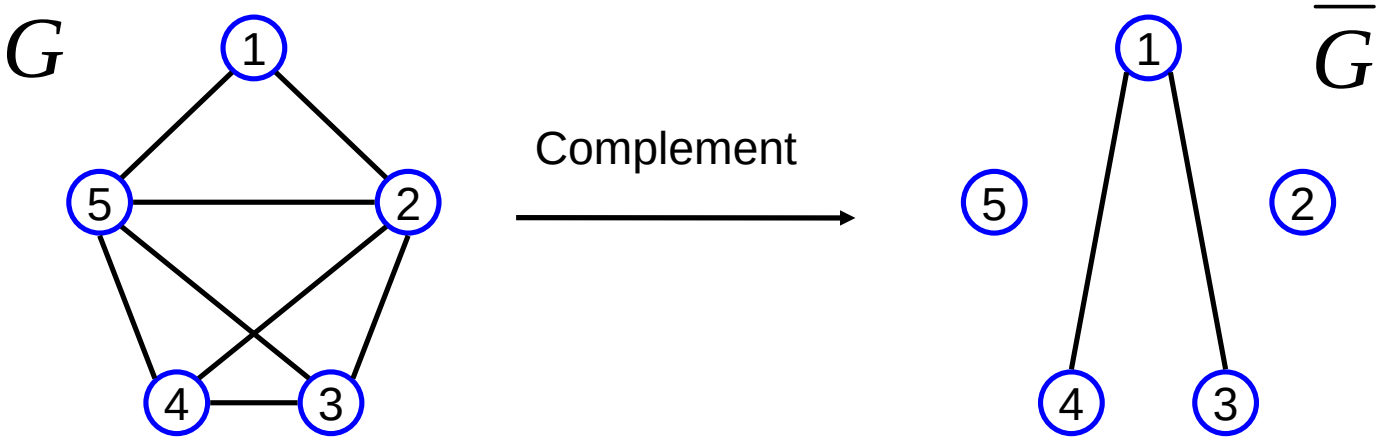
- actors are people and a tie exists if two people know each other.
- actors are wire transfer database records and a tie exists if two records have the same *matching field*.
- *Cohesive subgroups* are "closely knit groups" in a social network.
- *Social cohesion* is often used to explain and develop sociological theories.

Cliques

- *Etymology:* The term *clique* originates from Old French *cliquer* meaning *make a noise*
- *WordNet dictionary definition:* an exclusive circle of people with a common purpose
- Luce and Perry (1949): social clique – a group of people that know (are friends of) all other people in the group



Cliques and independent sets



{1,2,5} : maximal clique

{1,4} : maximal independent set

{2,3,4,5} : maximum clique

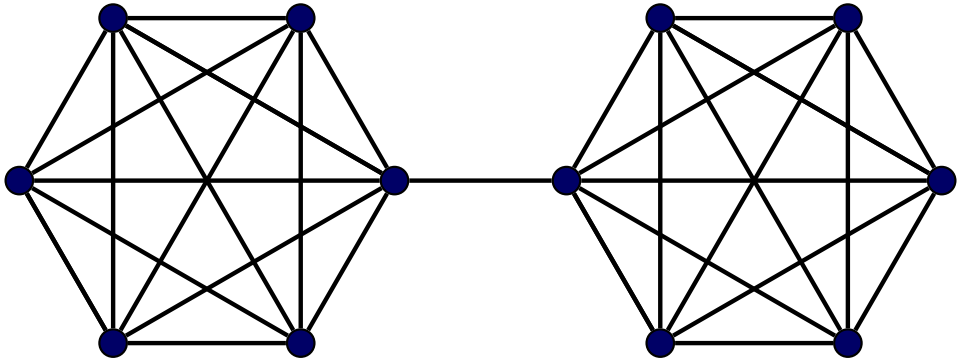
{1,2,5} : maximal independent set

{1,4} : maximal clique

{2,3,4,5} : maximum independent set

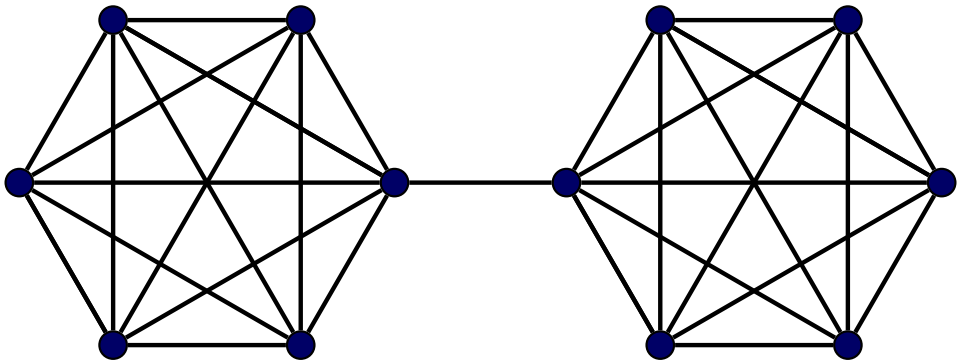
Clustering in networks

Partitioning vertices into 'natural groups' (clusters)



Clustering in networks

Partitioning vertices into 'natural groups' (clusters)



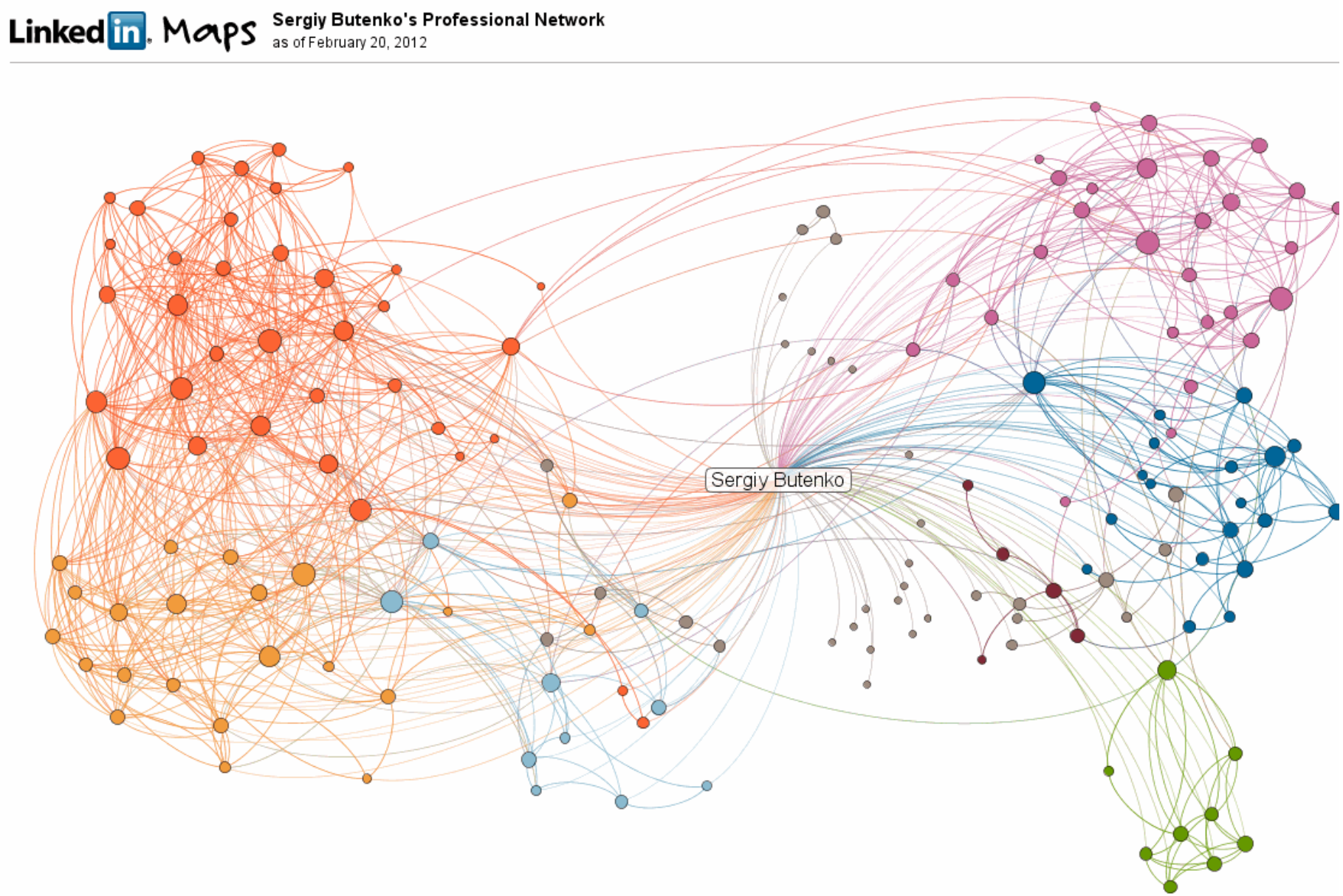
Community detection in graphs

S Fortunato - Physics reports, 2010 - Elsevier

The modern science of networks has brought significant advances to our understanding of complex systems. One of the most relevant features of graphs representing real systems is community structure, or clustering, ie the organization of vertices in clusters, with many ...

☆ ❞ Cited by 7012 Related articles All 48 versions »

Clustering in networks



Clustering in networks

Alternative approaches to unsupervised clustering:

- Require each cluster to satisfy certain structural properties and minimize the number of clusters

  Minimum clique partitioning

- Optimize a certain quantitative measure of clustering quality

  Modularity: the fraction of edges that fall within clusters, minus the expected fraction of edges within clusters for a random graph with same degree distribution as the given network.

Outline

1. Introduction
2. Clique relaxations
3. Asymptotic bounds in random networks
4. Conclusion

Alternatives to clique

$G = (V, E)$. $S \subseteq V$ is

- s-**clique** if $d_G(v, v') \leq s$, for any $v, v' \in S$ (Luce 1950)
- s-**club** if $diam(G[S]) \leq s$ (Alba 1973, Mokken 1979)
- s-**plex** if $\delta(G[S]) \geq |S| - s$ (Seidman & Foster 1978)
- s-**defective clique** if $G[S]$ has at least $\binom{|S|}{2} - s$ edges (Yu et al. 2006)
- k-**core** if $\delta(G[S]) \geq k$ (Seidman 1983)
- k-**block** if $\kappa(G[S]) \geq k$ (Moody & White 2003)
- γ-**quasi-clique** if $\rho(G[S]) \geq \gamma$ (Abello et al. 2002)
- (λ, γ)-**quasi-clique** if $\delta(G[S]) \geq \lambda(|S| - 1)$ and $\rho(G[S]) \geq \gamma$ (Brunato et al. 2008)
- ...

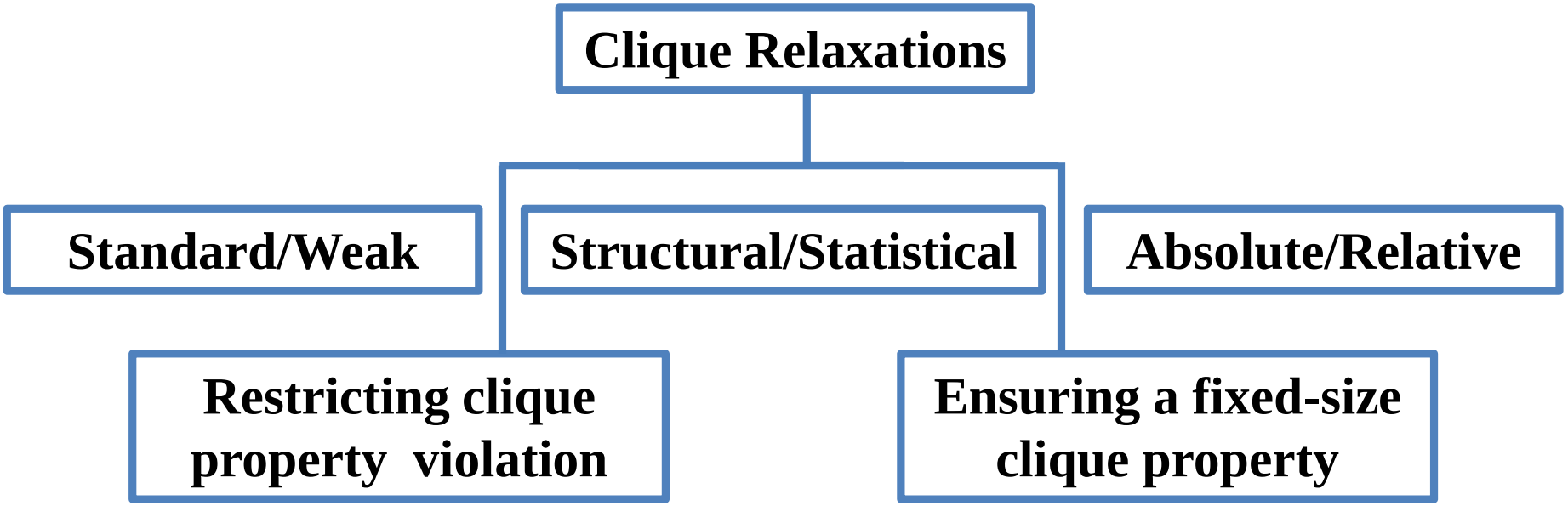
Elementary clique-defining properties

Proposition

A subset of vertices C is a clique in G if and only if one of the following conditions hold:

a) *Pairwise distances:* $d_G(v, v') = 1$, *for any* $v, v' \in C$;
b) *Diameter:* $diam(G[C]) = 1$;
c) *Domination:* $D = \{v\}$ *is a dominating set in* $G[C]$, *for any* $v \in C$;
d) *Minimum degree:* $\delta(G[C]) = |C| - 1$;
e) *Edge density:* $\rho(G[C]) = 1$;
f) *Vertex connectivity:* $\kappa(G[C]) = |C| - 1$.

Clique relaxations taxonomy



"The whole is more than the sum of its parts."
–Aristotle (384-322 BC)

J. Pattillo, N. Youssef, and S. Butenko. On clique relaxation models in network analysis. *European Jour. of Oper. Res.*, 226: 9–18, 2013.

Clique relaxations

B. Balasundaram, S. Butenko, and I. Hicks. Clique relaxations in social network analysis: the maximum k-plex problem. *Operations Research*, 2011.

S. Trukhanov, C. Balasubramaniam, B. Balasundaram, and S. Butenko. Algorithms for detecting optimal hereditary structures in graphs, with application to clique relaxations. *Computational Optimization and Applications*, 2013.

J. Pattillo, A. Veremyev, S. Butenko, V. Boginski. On the maximum quasi-clique problem. *Discrete Applied Math*, 2013

V. Boginski, S. Butenko, O. Shirokikh, S. Trukhanov, and J. Gil-Lafuente. A network-based data mining approach to portfolio selection via weighted clique relaxations. *Annals of Operations Research*, 2014.

A. Buchanan, J. S. Sung, V. Boginski, S. Butenko. On connected dominating sets of restricted diameter. *European Jour. of Oper. Res.*, 236: 410–418, 2014.

J. Pattillo, Y. Wang, and S. Butenko. Approximating 2-cliques in unit disk graphs. *Discrete Applied Math*, 2014.

S. Shahinpour and S. Butenko.Algorithms for the maximum k-club problem in graphs.*J. of Combinatorial Optim*, 2013.

A. Veremyev, O. A. Prokopyev, S. Butenko, and E. L. Pasiliao. Exact MIP-based approaches for finding maximum quasi-cliques and dense subgraphs. *Computational Optimization and Applications*, 2017.

C. Balasubramaniam and S. Butenko. On robust clusters of minimum cardinality in networks. *Annals of Operations Research*, 2017.

O. Yezerska, S. Butenko, and V. L. Boginski. Detecting robust cliques in graphs subject to uncertain edge failures. *Annals of Operations Research*, 2018.

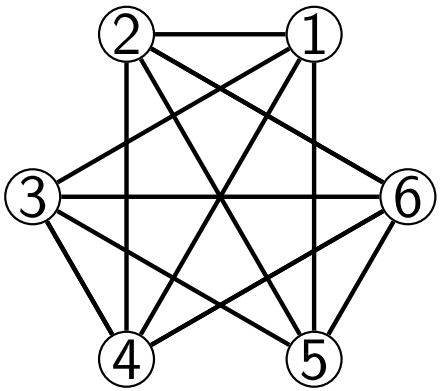
O. Yezerska, F. Mahdavi Pajouh, A. Veremyev, and S. Butenko. Exact algorithms for the minimum s-club partitioning problem. *Annals of Operations Research*, 2018.

Clique relaxations: s-defective clique

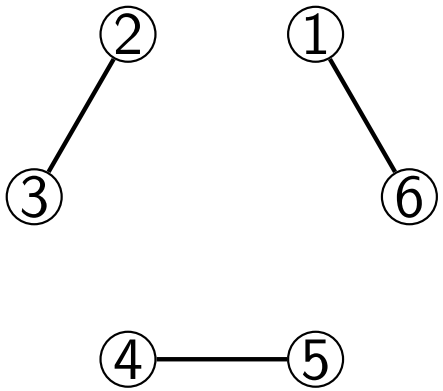
Definition

A subset of vertices S is an s-defective clique if the number of edges in the induced subgraph is at least $\binom{|S|}{2} - s$.

i.e., at most s edges are missing.



3-defective clique

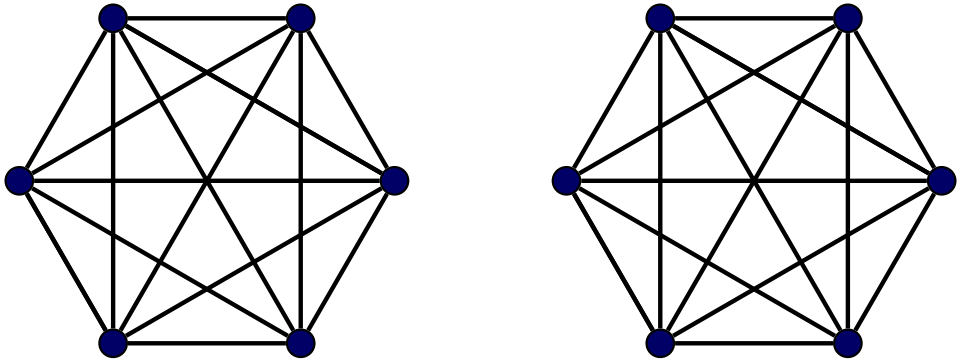


co-3-defective clique

Independent union of cliques

Definition (Independent union of cliques (IUC))

A subset of vertices C is called an independent union of cliques (IUC) if every connected component of $G[C]$ is a complete graph.



Z. Ertem, A. Veremyev, and S. Butenko. Detecting large cohesive subgroups with high clustering coefficients in social networks. *Social Networks* 46: 1–10, 2016.

Independent union of cliques

Definition

An *open triangle* is a simple graph with three vertices and two edges (i.e., a 3-vertex path), and a *closed triangle* is a complete 3-vertex graph.

Proposition

$C \subseteq V$ is an IUC *if and only if no set of three vertices from C induces an open triangle in G.*

Independent union of cliques

$$\alpha(G) = \max_{x \in [0,1]^n} \sum_{i \in V(G)} \frac{x_i}{1 + \sum_{j \in N(i)} x_j}$$

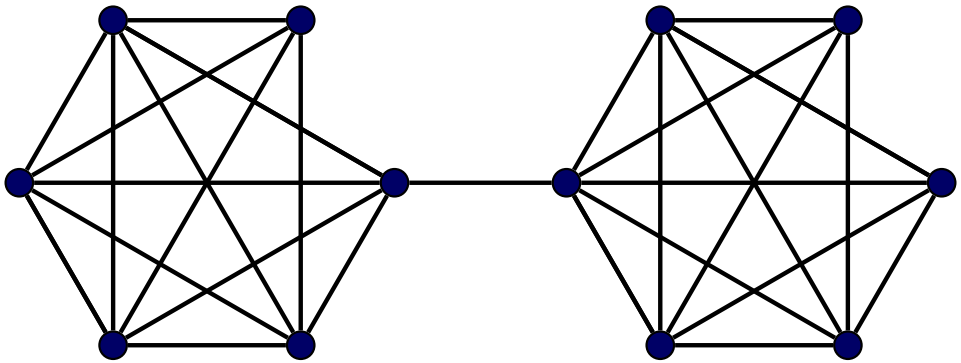
- If x^* is a global maximizer and $x^* \equiv I^*$, then I^* induces an IUC.

B. Balasundaram and S. Butenko. On a polynomial fractional formulation for independence number of a graph. *Journal of Global Optimization*, 2006.

Cluster editing

Cluster editing

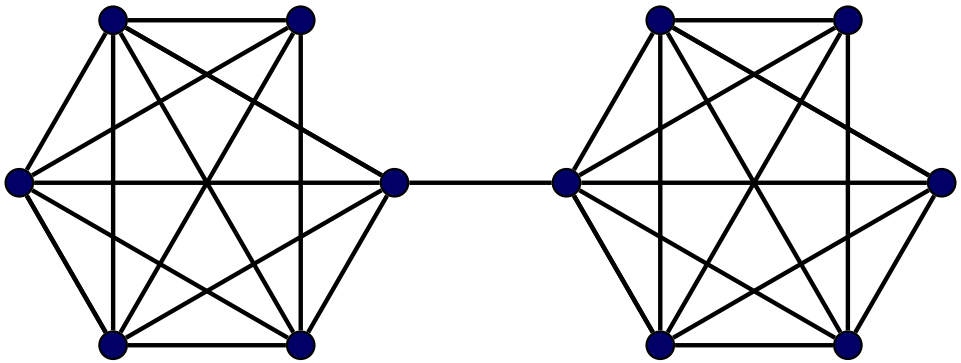
Remove or add minimum number of edges to obtain an IUC.



Cluster editing

Cluster editing

Remove or add minimum number of edges to obtain an IUC.



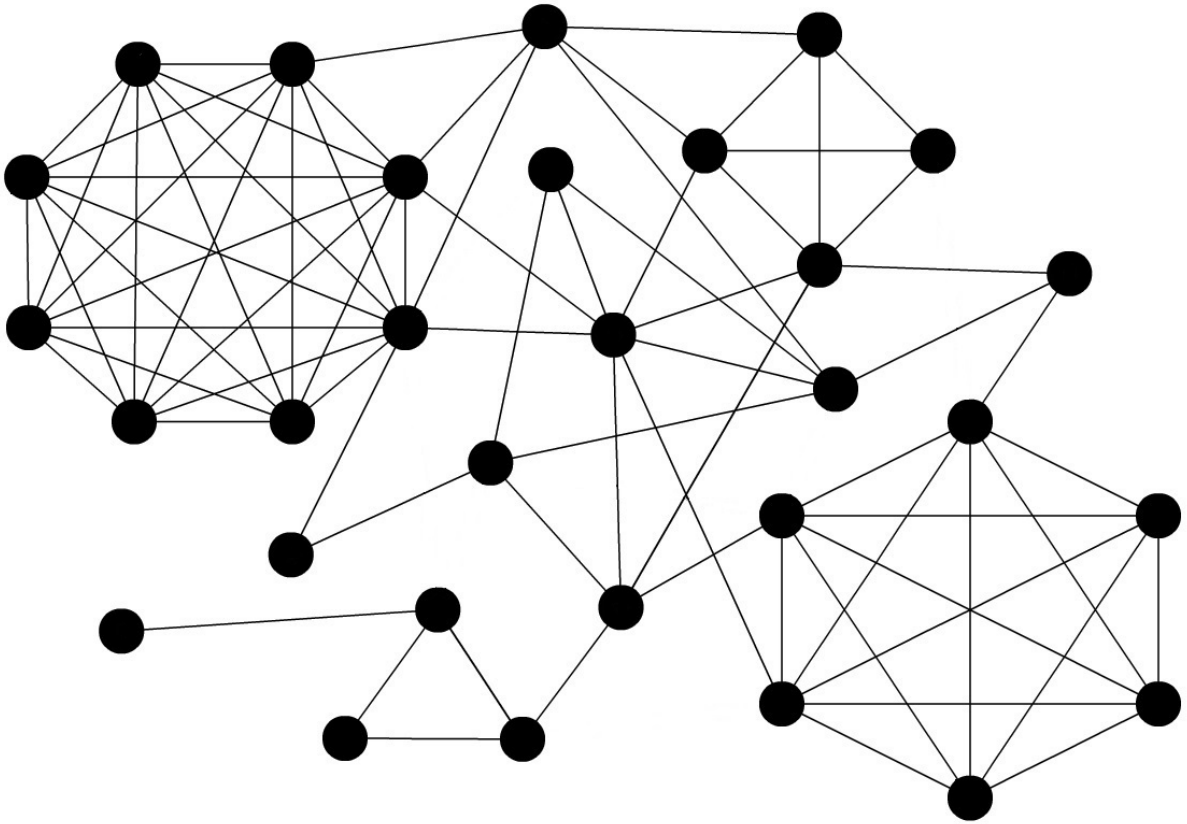
Cluster vertex deletion

Remove minimum number of vertices to obtain an IUC.

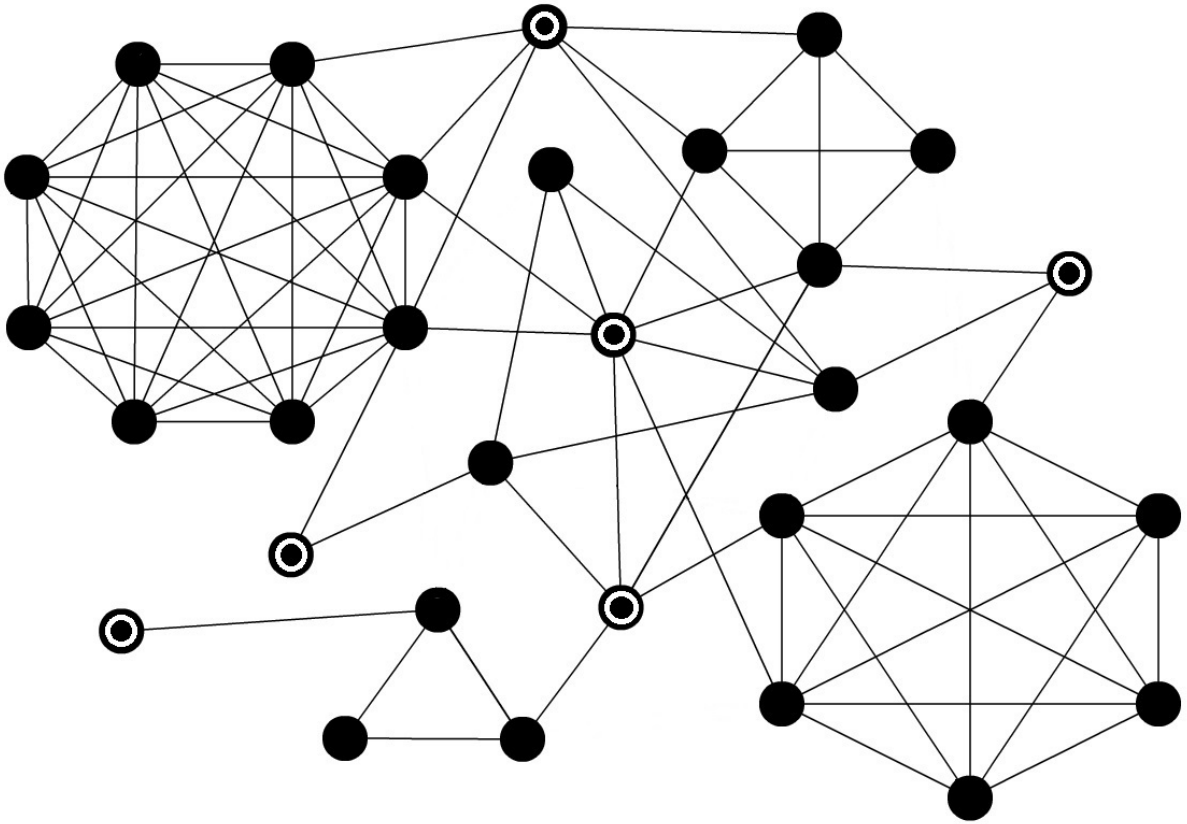
Independent union of cliques and cluster vertex deletion

- Maximum IUC is equivalent to the optimization version of *cluster vertex deletion problem* (Hüffner et al., 2010) (or the s-*plex cluster vertex deletion problem* with $s = 1$ (Bevern et al., 2012)
  - Find a minimum number of vertices that need to be removed from the graph so that the remaining vertices form an IUC.
  - D is an optimal solution of the cluster deletion problem if and only if $C = V \setminus D$ is a maximum IUC.

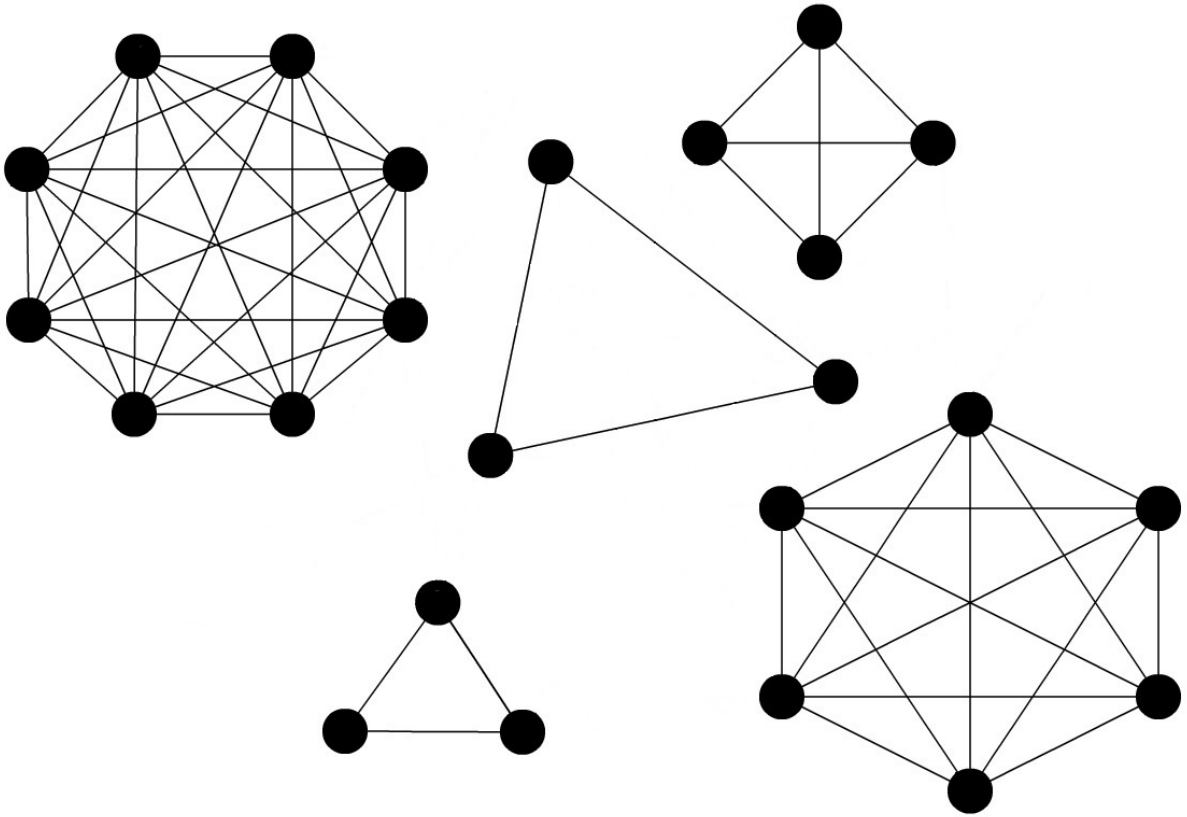
Independent union of cliques and cluster vertex deletion



Independent union of cliques and cluster vertex deletion



Independent union of cliques and cluster vertex deletion



cluster graph

Heredity in induced subgraphs

A graph property Π said to be *nontrivial*, if every graph with one vertex satisfies Π, and not every graph satisfies Π. A graph property Π said to be *interesting*, if there are graphs with arbitrary large number of vertices satisfying Π.

A graph property Π said to be *hereditary* on induced subgraphs, if for every $S \subseteq V$ such that $G[S]$ satisfies Π, the property Π is also satisfied by $G[S']$ for any nonempty $S' \subseteq S$.

Theorem (Yannakakis, 1978)

The maximum Π problem for a property Π that is nontrivial, interesting and hereditary on induced subgraphs is $\mathcal{NP}$-hard.

Heredity in induced subgraphs

- Clique
- Independent (stable) set
- Forest
- Bipartite graph
- Perfect graph
- s-defective clique
- s-plex
- s-bundle
- Independent union of cliques

Independent unions

Definition

Independent union of Π-sets ($IU(\Pi)$) – a subset of vertices inducing a subgraph where every connected component satisfies Π.

- $\Pi =$ "clique": independent union of cliques (IUC).
- $\Pi =$ "s-defective clique": independent union of s-defective cliques (IUsDC).

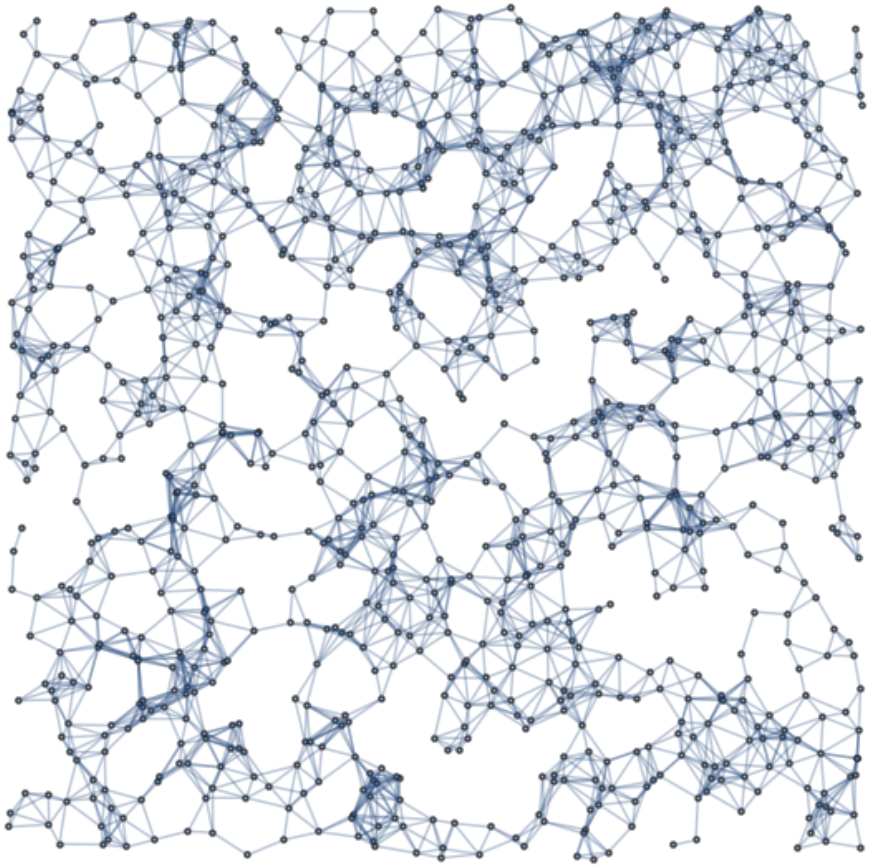
Obvervations

- if Π is hereditary, then $IU(\Pi)$ is hereditary
- if Π is nontrivial, then independent set satisfies $IU(\Pi)$

Outline

1. Introduction
2. Clique relaxations
3. Asymptotic bounds in random networks
4. Conclusion

Uniform random graphs



- Uniform random graph model (Erdős and Rényi): Given a set of n vertices V, each edge exists with probability $0 \leq p \leq 1$.
- Denoted $G(n, p)$.
- Testing scalability of algorithms.
- Is the problem easy "on average"?
- Can we bound the solution value (with high probability)?

Defining functions

Definition

A defining function $\zeta(\Pi; n, p)$ for property Π is the probability of a random graph $G(n, p)$ satisfying Π.

Π	$\zeta(\Pi; n, p)$
Clique	$p^{\binom{n}{2}}$
Stable set	$(1-p)^{\binom{n}{2}}$
s-defective clique	$p^{\binom{n}{2}} + \cdots + \binom{\binom{n}{2}}{s} p^{\binom{n}{2}-s}(1-p)^s$
γ-quasi clique	$p^{\binom{n}{2}} + \cdots + \binom{\binom{n}{2}}{\lceil \gamma \binom{n}{2} \rceil} p^{\lceil \gamma \binom{n}{2} \rceil}(1-p)^{\binom{n}{2} - \lceil \gamma \binom{n}{2} \rceil}$

Defining functions

Basic properties

- if Π is nontrivial, then $\zeta(\Pi; 1, p) = 1$
- if Π is hereditary, then $\zeta(\Pi; n, p) \leq \zeta(\Pi; n-1, p)$
- $\zeta(\Pi; n, p) \leq \zeta(IU(\Pi); n, p)$

Classical approach

Question

What is the cardinality of a maximum Π-set in $G(n,p)$ with high probability?

- Consider a random variable

$$X_k = \text{number of } \Pi\text{-sets of cardinality } k \text{ in } G(n,p).$$

Then

$$\mathbb{P}\left(X_k \geq 1\right) \leq \mathbb{E}[X_k] = \binom{n}{k} \zeta(\Pi; k, p)$$

- We are interested in the largest k, such that $\mathbb{E}[X_k] \geq 1$.
- If $\mathbb{E}[X_{f(n)}] \to 0$, $G(n,p)$ does not contain a Π-set of cardinality $f(n)$ w.h.p.

Lower bound

Lemma

For a nontrivial property Π, the cardinality of the maximum $IU(\Pi)$-set in $G(n,p)$ is at least $(1+\epsilon)\log_{\frac{1}{1-p}}(n)$ w.h.p.

Lemma

For a nontrivial property Π, if clique satisfies Π, then the cardinality of the maximum $IU(\Pi)$-set in $G(n,p)$ is at least $(1+\epsilon)\log_{\frac{1}{p}}(n)$ w.h.p.

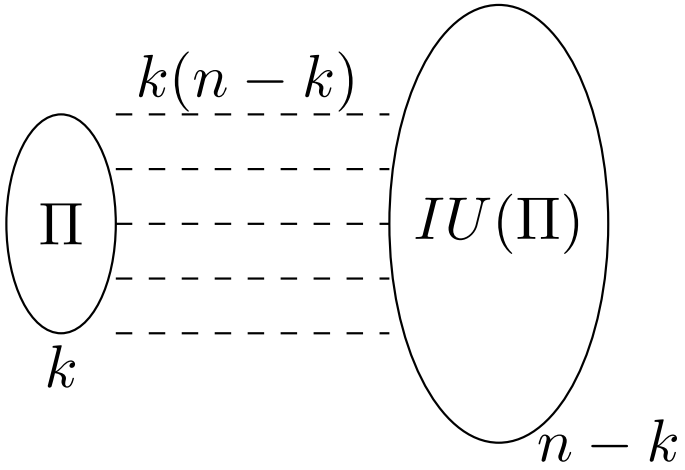
Defining function of independent unions

Theorem (IU defining function recurrence)

Let a nontrivial, interesting property Π have a defining function $\zeta(\Pi; n, p)$. Then

$$\zeta(IU(\Pi); n, p) = \sum_{k=1}^{n} \binom{n-1}{k-1} \zeta(\Pi; k, p)(1-p)^{k(n-k)} \zeta(IU(\Pi); n-k, p),$$

where $\zeta(IU(\Pi); 0, p) = \zeta(IU(\Pi); 1, p) = 1$.

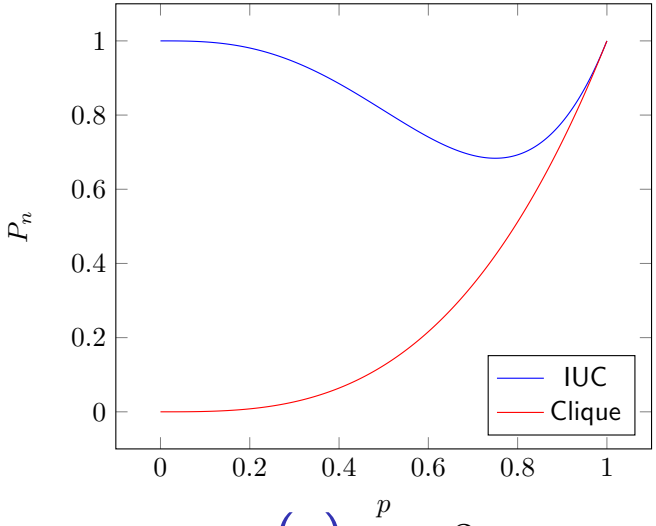


IUC defining function

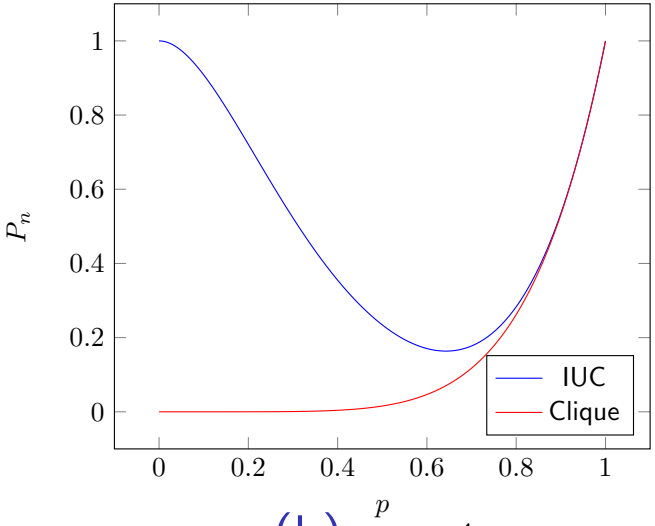
Recall, for Π="clique", $\zeta(\Pi; n, p) = p^{\binom{n}{2}}$. For Π="IUC", we have:

n	$\zeta(\Pi; n, p)$
0-2	1
3	$1 - 3p^2(1-p)$
4	$1 - 12p^2 + 32p^3 - 39p^4 + 24p^5 - 5p^6$
5	$12p^{10} - 50p^9 + 75p^8 + 10p^7 - 195p^6 +$
	$+318p^5 - 270p^4 + 130p^3 - 30p^2 + 1$
6	$1 - 60p^2 + 360p^3 - 945p^4 + 792p^5 + 2775p^6 - 12090p^7$
	$+25035p^8 - 34240p^9 + 33204p^{10} - 23130p^{11} + 11385p^{12} -$
	$-3780p^{13} + 765p^{14} - 71p^{15}$

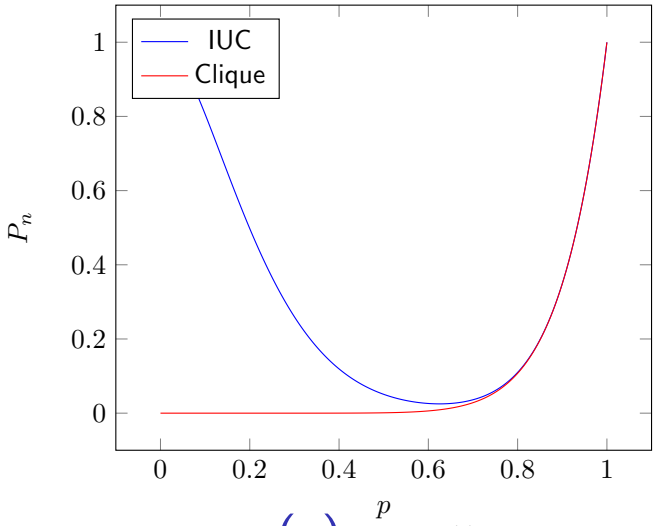
IUC defining function



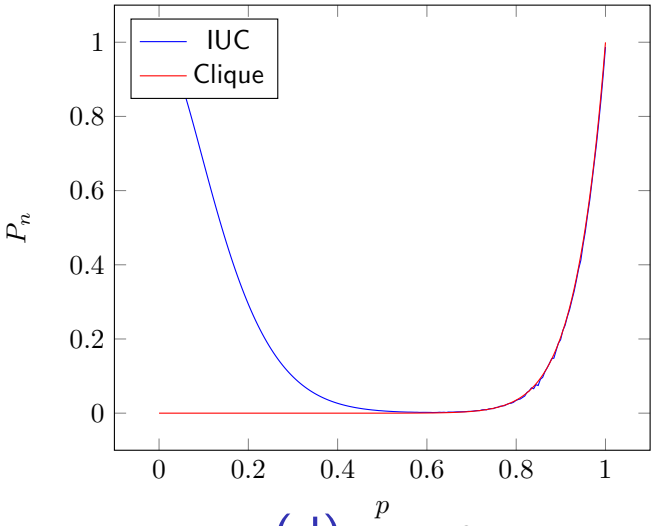
(a) $n = 3$



(b) $n = 4$



(c) $n = 5$



(d) $n = 6$

Conditions for logarithmic upper bounds

Theorem

If $\frac{\zeta(\Pi;n+1,p)}{\zeta(\Pi;n,p)} = O(np^n)$ *for fixed* $p \geq 0.5$*, then* $\max |IU(\Pi)| \leq 2\left\lceil \log_{\frac{1}{p}}(n) \right\rceil + 1$ *with high probability.*

Theorem

If $\frac{\zeta(\Pi;n+1,p)}{\zeta(\Pi;n,p)} = O(n(1-p)^n)$ *for fixed* $p \leq 0.5$ *then* $\max |IU(\Pi)| \leq 2\left\lceil \log_{\frac{1}{1-p}}(n) \right\rceil + 1$ *with high probability.*

Lemma

The results of the previous theorems hold even when $\zeta(\Pi, n, p)$ *is replaced with some* $\hat{\zeta}(\Pi, n, p) \geq \zeta(\Pi, n, p)$*.*

IUC in $G(n,p)$: $p > 0.5$

Theorem

Let $\alpha^\omega(G)$ be the cardinality of a maximum IUC in G. Then for $p > 0.5$, we have

$$\alpha^\omega(G(n,p)) \sim \omega(G(n,p)).$$

This implies that

$$\alpha^\omega(G(n,p)) \leq 2\log_{\frac{1}{p}}(n)$$

with high probability.

IUC in $G(n,p)$: $p < 0.5$

Theorem

Let $\alpha^\omega(G)$ be the cardinality of a maximum IUC in G. Then for $p > 0.5$, we have

$$\alpha^\omega(G(n,p)) \sim \alpha(G(n,p)).$$

This implies that

$$\alpha^\omega(G(n,p)) \leq 2\log_{\frac{1}{1-p}}(n)$$

with high probability.

IUC in $G(n,p) : p = 0.5$

- Recall that $\zeta(clique; n, p) = p^{\binom{n}{2}}$.
  - If $p = 0.5$, then $\zeta(clique; n, 0.5) = 2^{-\binom{n}{2}}$.

Theorem

Let B_n be the n-th Bell number (the number of partitions of a set of size n). Then

$$\zeta(IUC; n, p) = 2^{-\binom{n}{2}} \cdot B_n.$$

Moreover,

$$\alpha^{\omega}(G(n, 0.5)) \leq 2\log_2(n)$$

with high probability.

IUC in $G(n,p) : p = 0.5$

Proof:

- According to the IU defining function recurrence,

$$\zeta(IUC; n, 0.5) = 2^{-\binom{n}{2}} \sum_{t=0}^{n-1} \binom{n-1}{t} 2^{\binom{t}{2}} \zeta(IUC; t, 0.5).$$

- Define $B_n = 2^{\binom{n}{2}} \zeta(IUC; n, 0.5)$. Then B_n satisfies the recursive relation

$$B_n = \sum_{t=0}^{n-1} \binom{n-1}{t} B_t \text{ with } B_0 = 1.$$

- This relation is precisely the relation for the n-th Bell number. □

IUC in $G(n,p) : p = 0.5$

- Let X_k = number of IUC of cardinality k, then

$$\mathbb{P}\left(X_k \geq 1\right) \leq \mathbb{E}[X_k] = \binom{n}{k} 2^{-\binom{k}{2}} \cdot B_k.$$

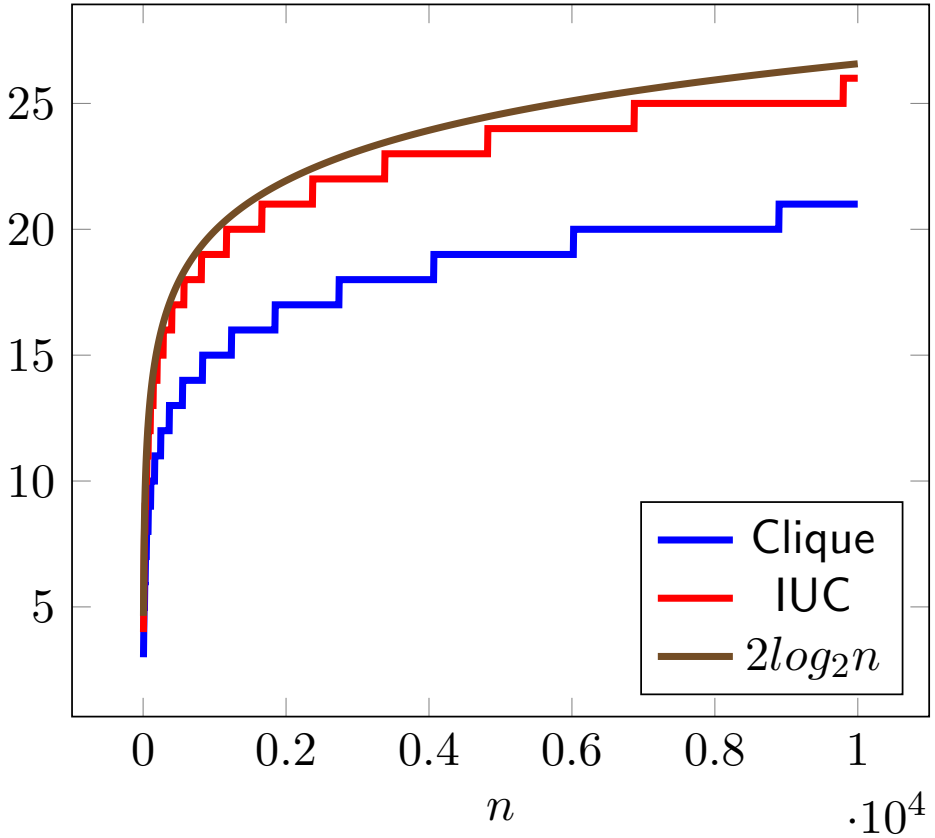
- According to Berend and Tassa (2010),

$$B_k \leq \left(\frac{0.792k}{\ln(k+1)}\right)^k.$$

- Thus,

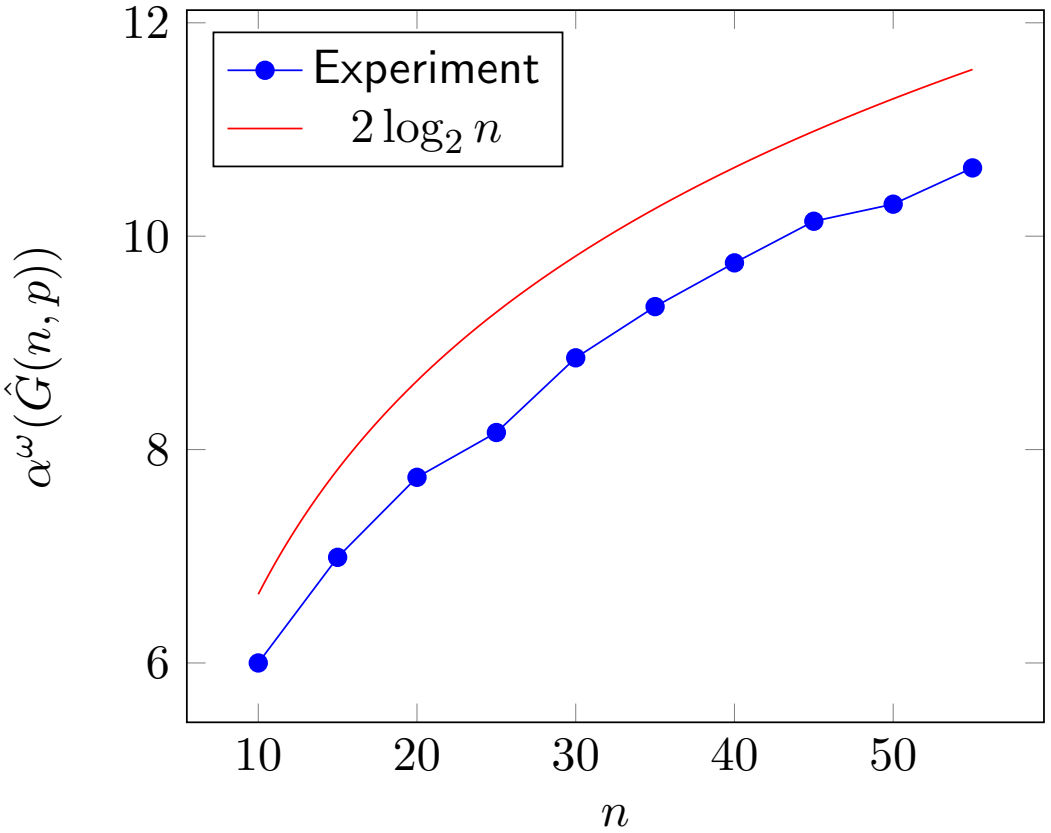
$$\mathbb{E}(X_k) = \binom{n}{k} 2^{-\binom{k}{2}} B_k \leq \binom{n}{k} 2^{-\binom{k}{2}} \left(\frac{0.792k}{\ln(k+1)}\right)^k.$$

IUC in $G(n,p) : p = 0.5$



Maximum k for which the theoretical bound on $\mathbb{E}(X_k)$ is at least 1.

IUC in $G(n,p) : p = 0.5$



s-Defective clique

- The defining function is

$$\zeta(n) = p^{\binom{n}{2}} + \sum_{i=1}^{s} \binom{\binom{n}{2}}{i} p^{\binom{n}{2}-i}(1-p)^i$$

- We have

$$\zeta(\Pi; n, p) \leq \hat{\zeta}(\Pi; n, p) = \binom{\binom{n}{2}}{s} p^{\binom{n}{2}-s}$$

and

$$\frac{\hat{\zeta}(n+1)}{\hat{\zeta}(n)} \leq 2^s n p^n.$$

- So, the logarithmic bound holds.

Outline

1. Introduction
2. Clique relaxations
3. Asymptotic bounds in random networks
4. Conclusion

Conclusions

- Criteria when independent unions have a logarithmic upper bound.
- Logarithmic upper and lower bound for the maximum IUC size in a uniform random graph.

Future work:

- Power-law graphs.
- Apply results to other independent unions, e.g., γ-quasi cliques and other clique relaxations.

thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
sukriya
kop khun krap
grazie
arigatô
takk
dakujem
merci
спасибо
рахмат
Баярлалаа
dank je
tapadh leat
mauruuru
hvala
mochchakkeram
terima kasih
감사합니다
ありがとう
ευχαριστώ
xièxie
merci
merci
merci