Dense subgraphs with hierarchical relations: Models, Algorithms, Applications

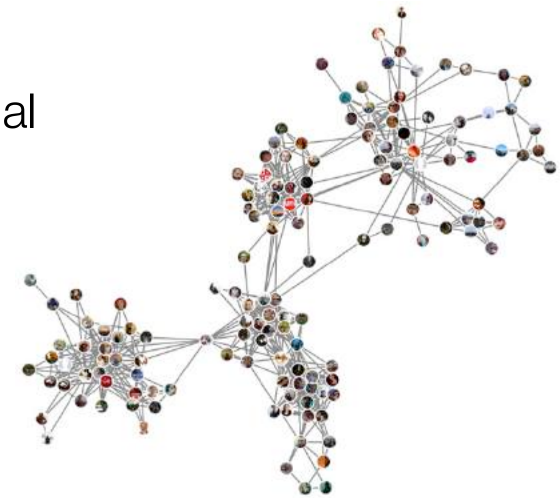
A. Erdem Sariyuce

University at Buffalo

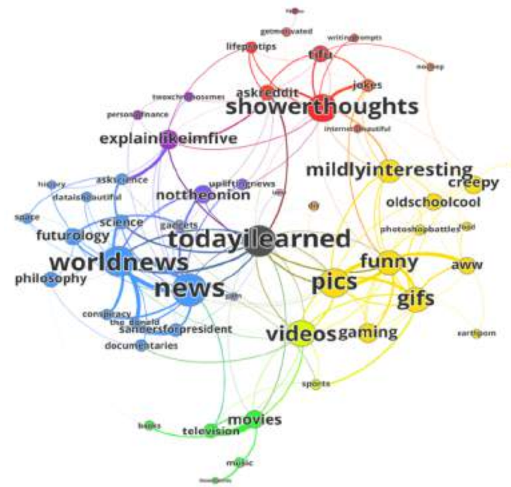
http://sariyuce.com

Graphs all around

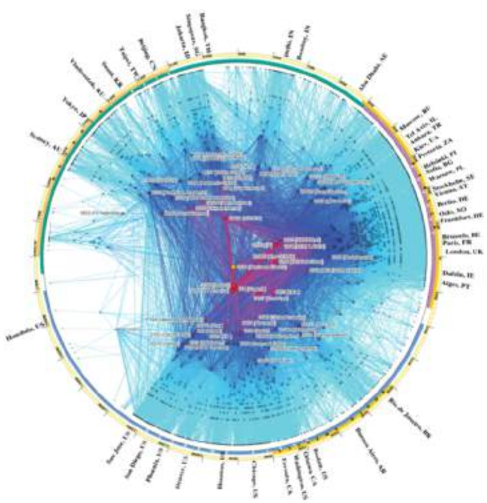
Social



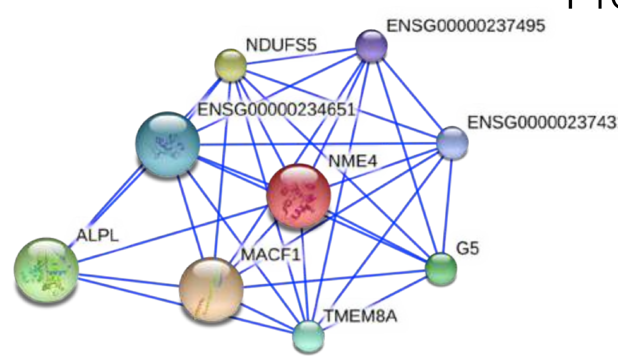
Information



Routers

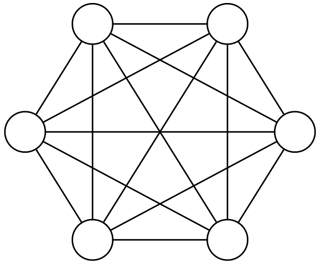


Protein-interaction



Dense subgraph discovery

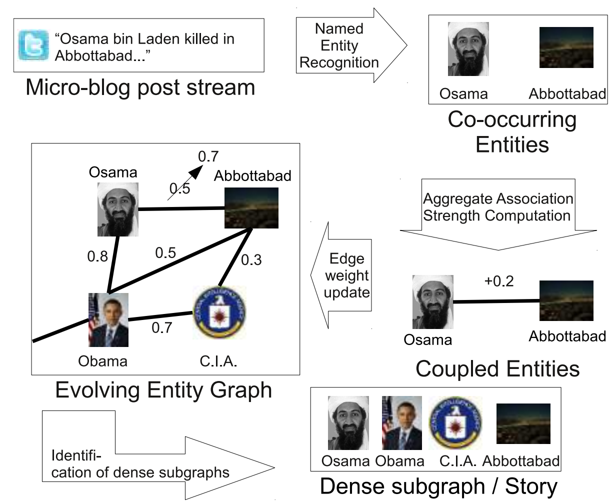
- Measure of connectedness on edges
  - # edge / # all possible
    - |E| / (|V| choose 2), 1.0 for a clique

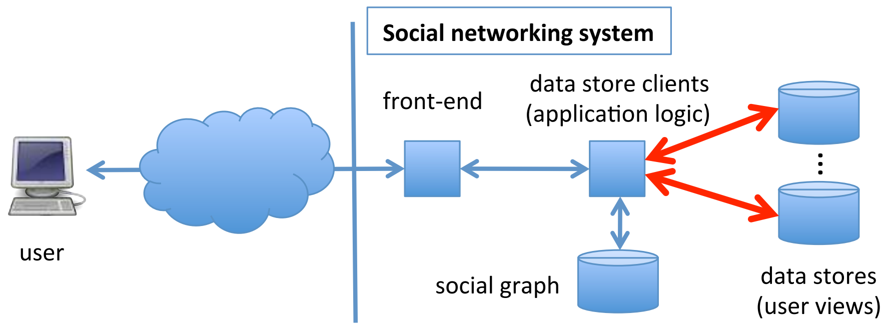


- Globally sparse, locally dense
  - $|E| << |V|^2$, but vertex neighborhoods are dense
    - High clustering coefficients – density of neighbor graph
- Many nontrivial subgraphs with high density
  - And relations among them
- Not clustering: Absolute vs. relative density

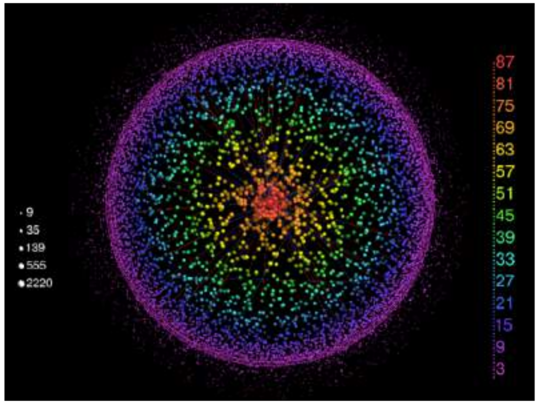
Dense subgraphs matter in many applications

- Significance or anomaly
  - Spam link farms [Gibson et al., '05]
  - Real-time stories [Angel et al., '12]



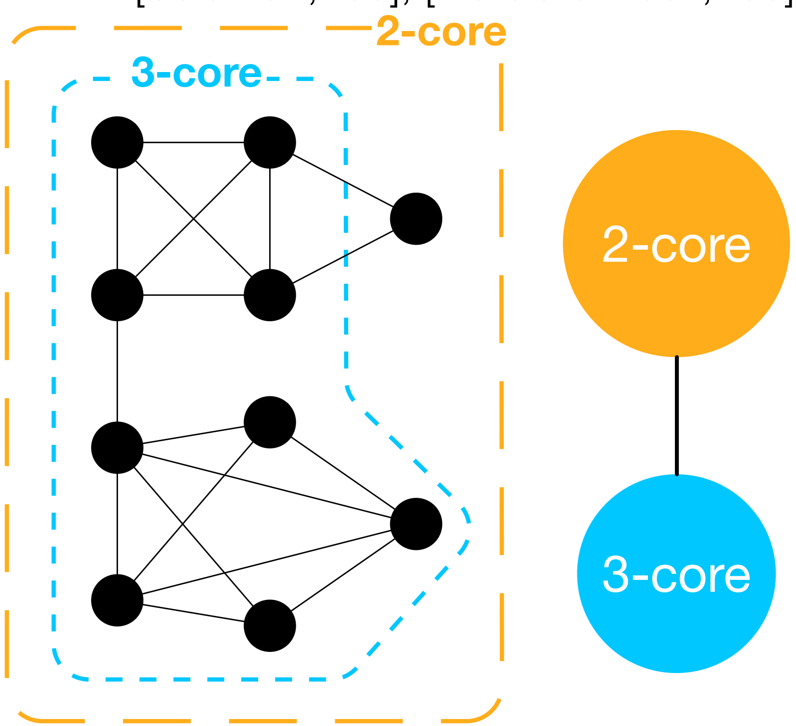


- Computation & summarization
  - System throughputs [Gionis et al., '13]
  - Graph visualization [Alvarez et al., '06]

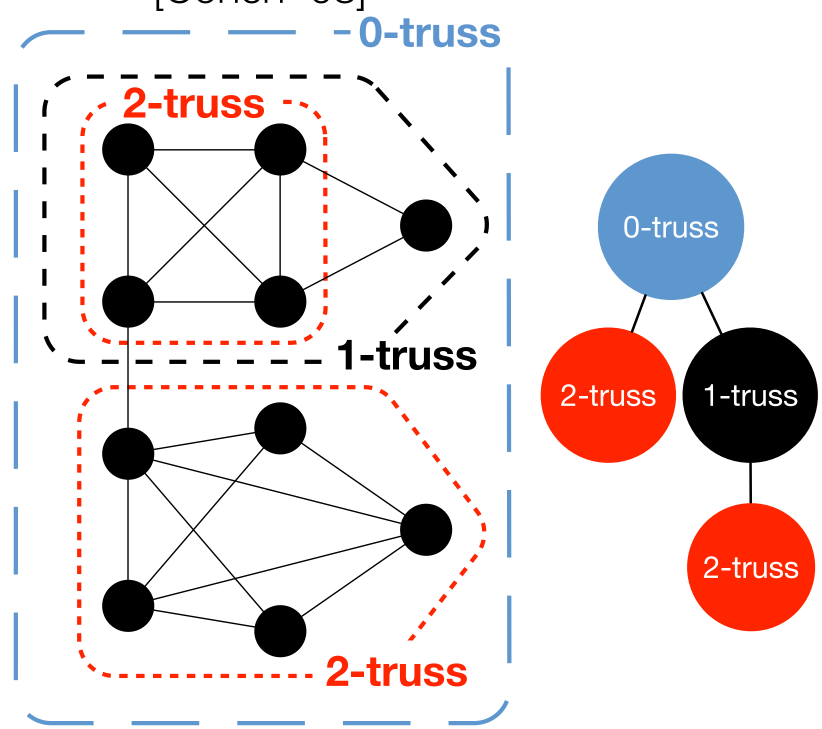


Two effective algorithms to find dense subgraphs with hierarchical relations

- ***k*-core:** Every vertex has at least *k* edges
  - [Seidman, ‘83], [Matula & Beck, ‘83]

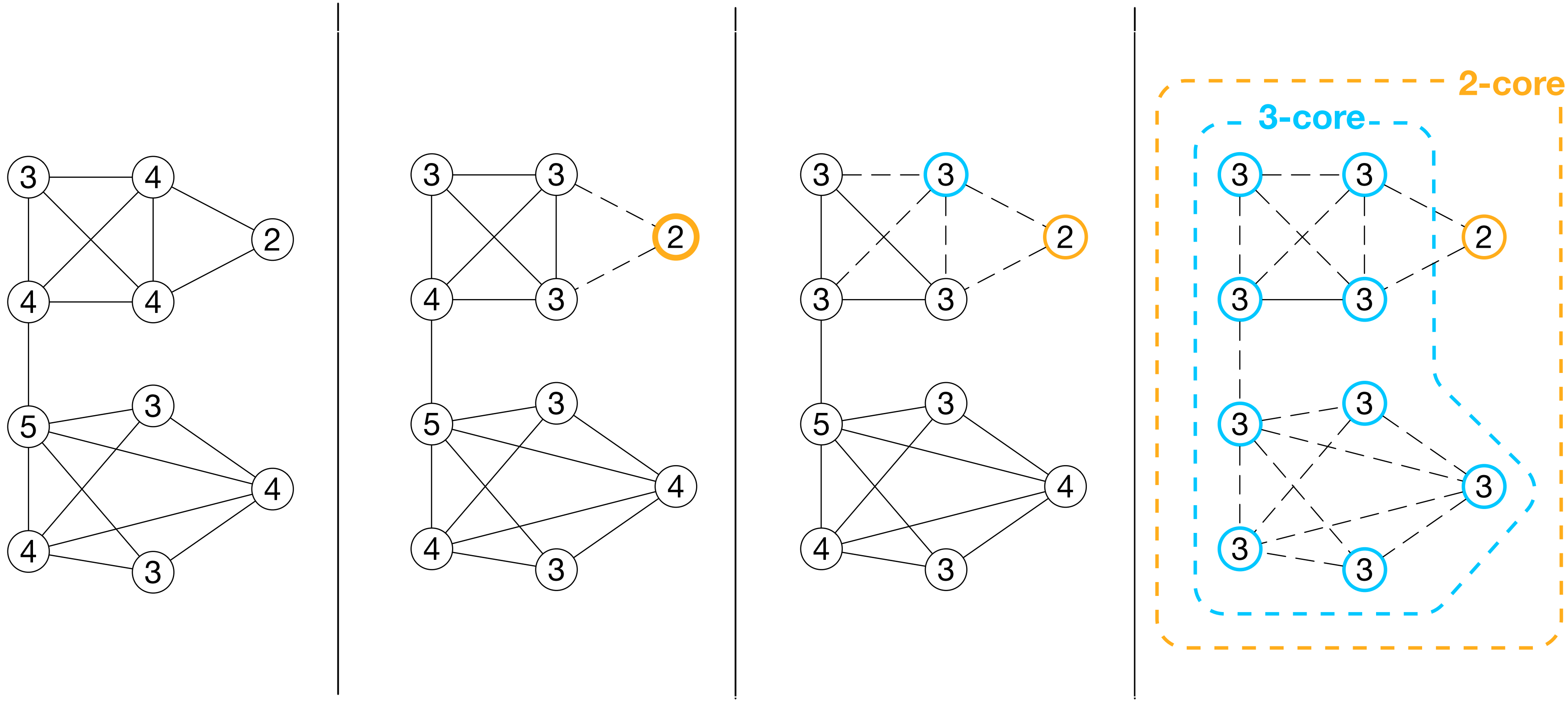


- ***k*-truss:** Every edge has at least *k* triangles
  - [Cohen ‘08]



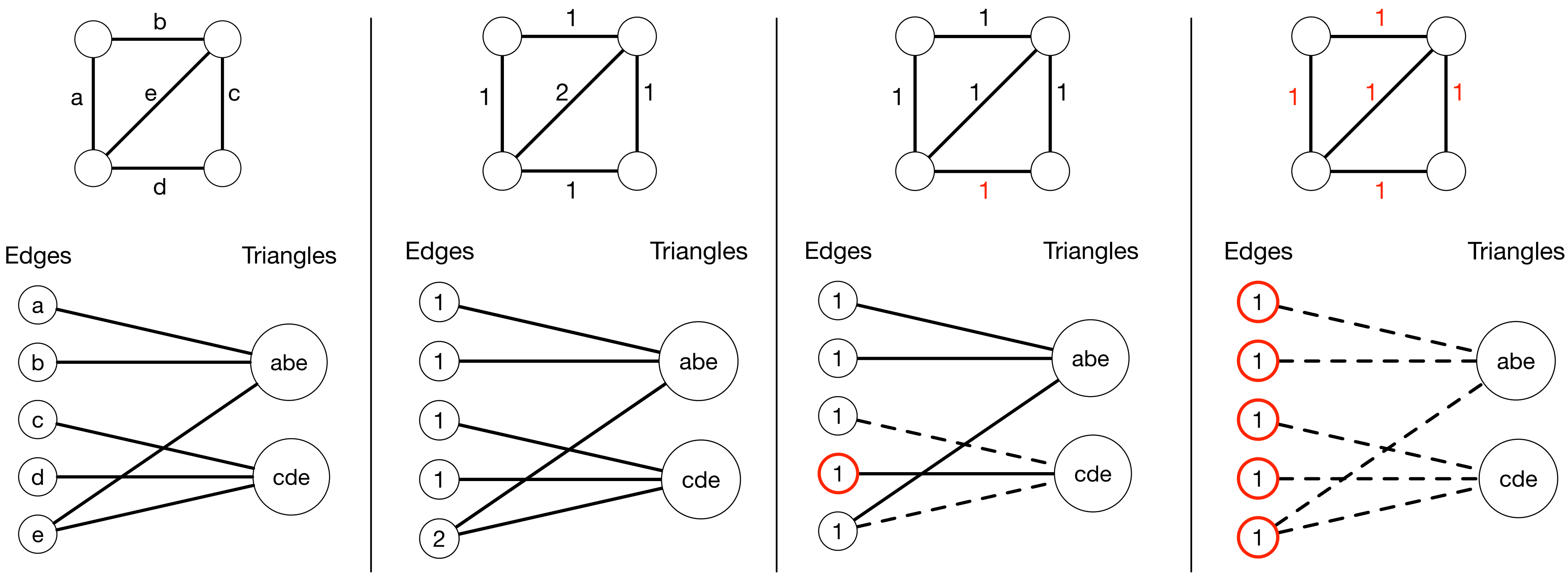
Peeling algorithm finds the *k*-cores & *k*-trusses

- Core numbers of vertices. O(|E|) [Matula & Beck, '83]
- Truss numbers of edges. O(|△|) [Cohen '08]



Observation: *k*-truss IS just *k*-core on the edge-triangle graph!

- Edge and triangle relations
  - Not a binary relation – three edges in a triangle
- Build a bipartite graph!



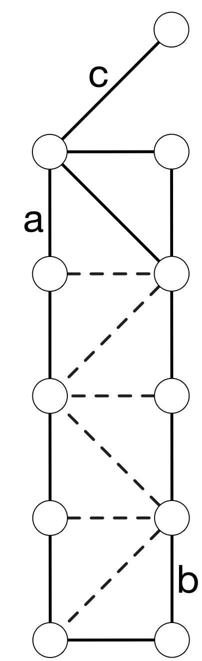
Why limit to triangles?

- Small cliques in larger cliques
  - 1-cliques in 2-cliques (vertices and edges)
  - 2-cliques in 3-cliques (edges and triangles)
- Generalize for any clique
  - r-cliques in s-cliques ($r < s$)
- Convert to bipartite
  - r-cliques → left vertices
  - s-cliques → right vertices
  - Connect if right contains left

Nucleus decomposition generalizes *k*-core and *k*-truss algorithms

- Say R is r-clique, S is s-clique ($r < s$)
- k-(r, s) nucleus: Every R takes part in at least k number of S
  - Each R_i, R_j pair is connected by series of Ss

***r*=1, *s*=2**	***r*=2, *s*=3**
***k*-core**	***k*-truss**
	(stronger conn.)
d(v) ≥ k	△(e) ≥ k
Simply connected	Triangle connected

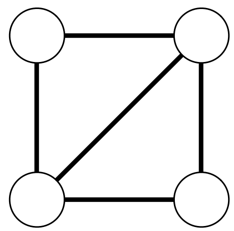


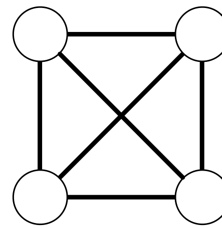
Sariyuce, Seshadhri, Pinar, Catalyurek, Int. WWW Conf., 2015 (Best paper runner-up)

Some nucleus examples

1 - 1 3 nucleus

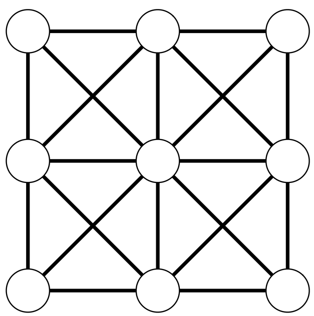
Each vertex has at least
one triangle





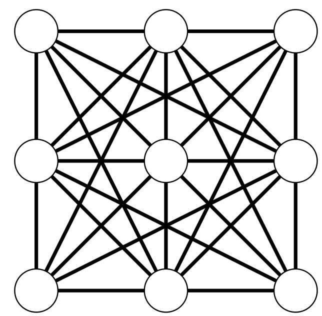
1 - 1 4 nucleus

Each vertex has at least
one 4-clique



2 - 2 3 nucleus

Each edge has at least
two triangles

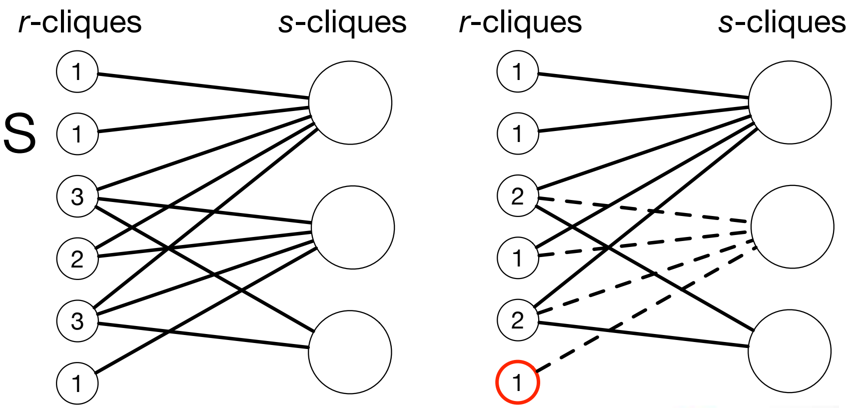


2 - 2 4 nucleus

Each edge has at least
two 4-cliques

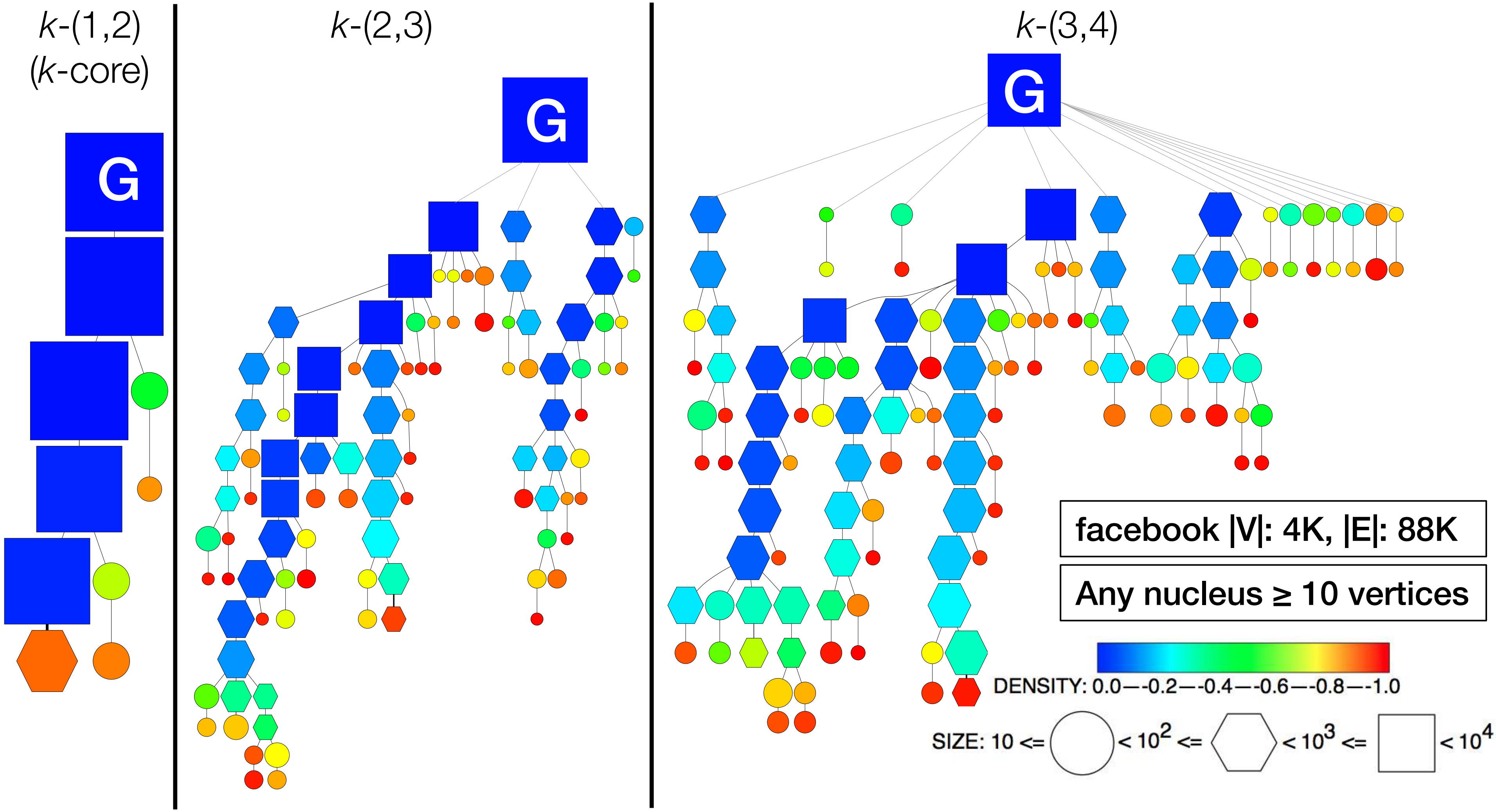
Peeling works for nucleus decomposition as well!

- On the bipartite graph
  - For vertex set of *r*-cliques
  - Degree based



- Sounds expensive?
  - Yes, in theory
  - *r*=3, *s*=4: $O(\sum_v cc(v)d(v)^3)$
  - But practical
    - Clustering coefficients decay with the degree in many real-world networks
  - Can be scaled to tens of millions of edges

Comparing hierarchies for different nucleus decompositions



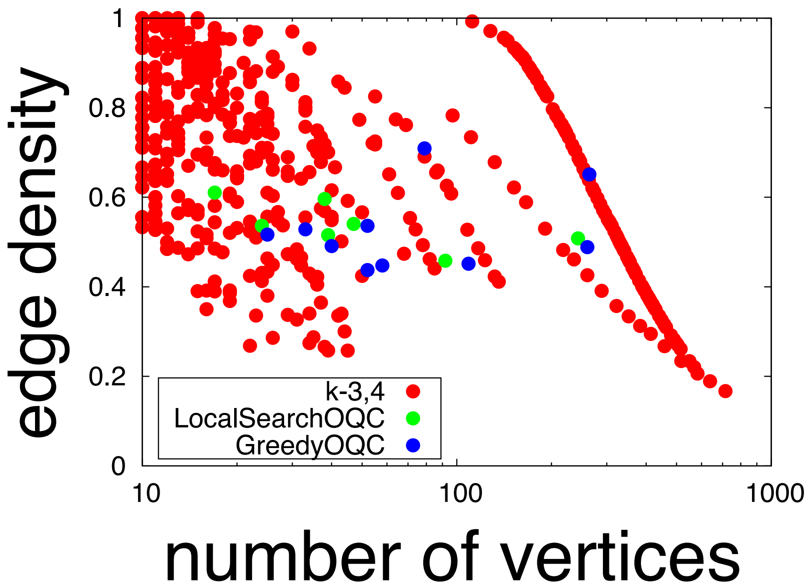
Comparison with the state-of-the-art

- Optimal quasi-cliques [Tsourakakis et al., 2013]
  - Almost-cliques, finds *only one* dense structure
- We find *all* subgraphs & relations -- *k*-(3,4)

---

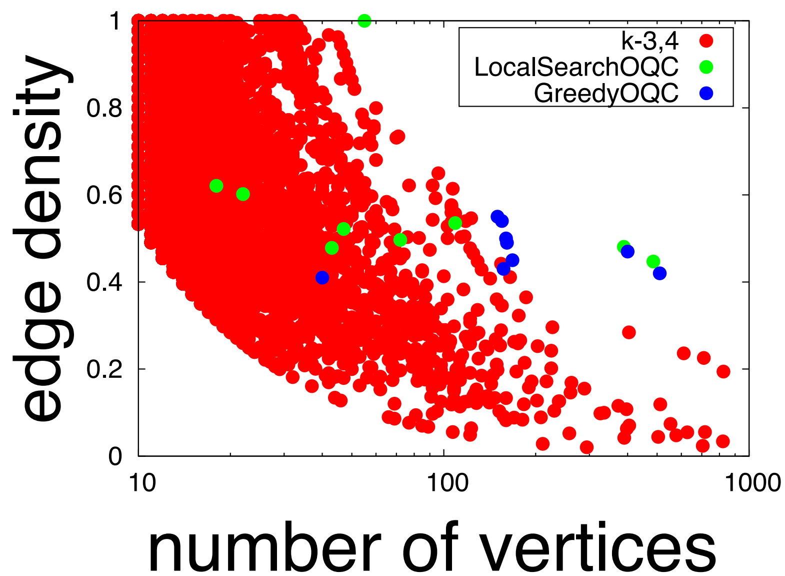
wikipedia-200611

|V|: 3.1M |E|: 39.3M

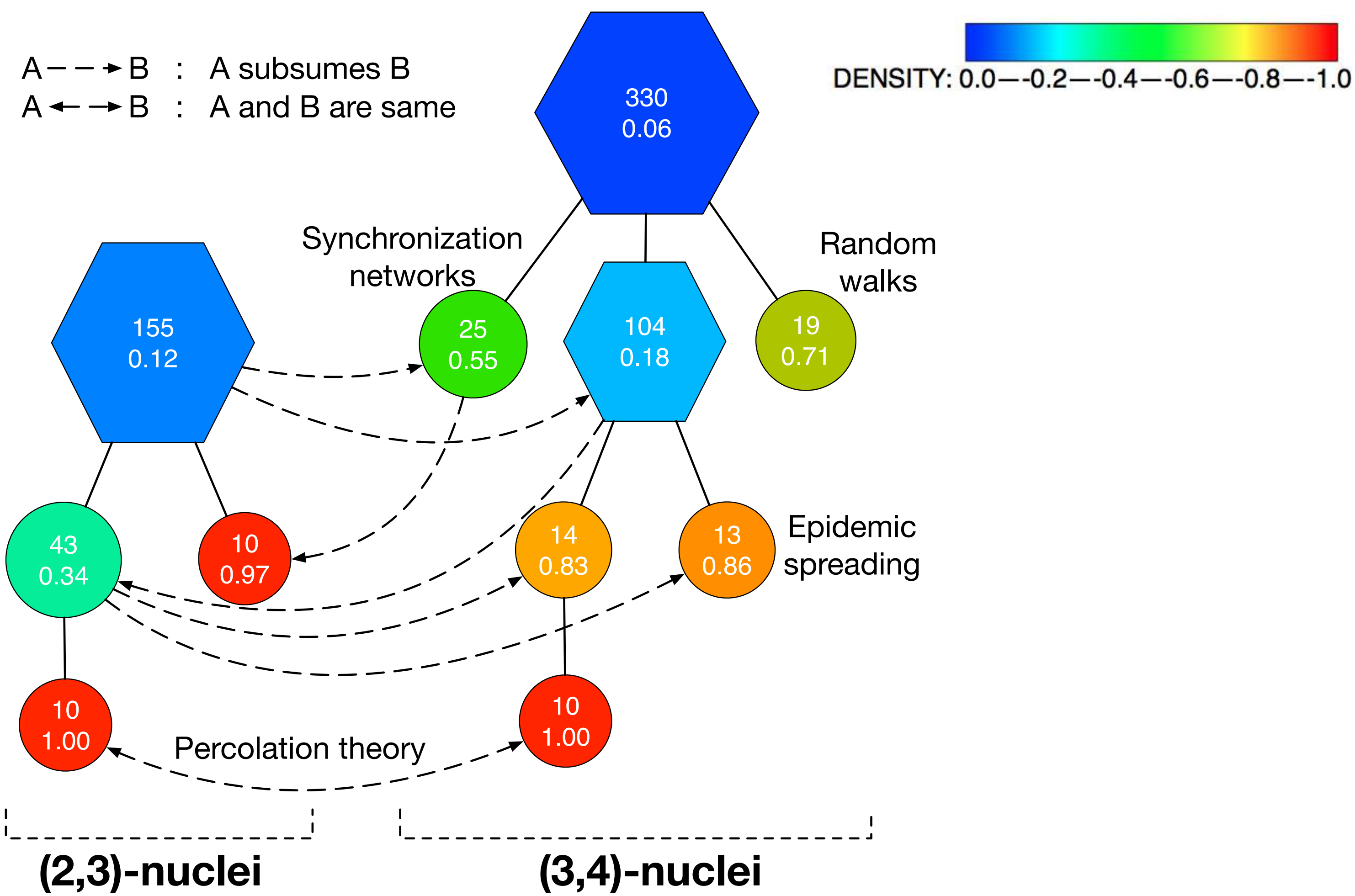


soc-sign-epinions

|V|: 131K |E|: 841K



APS Citation Network Analysis

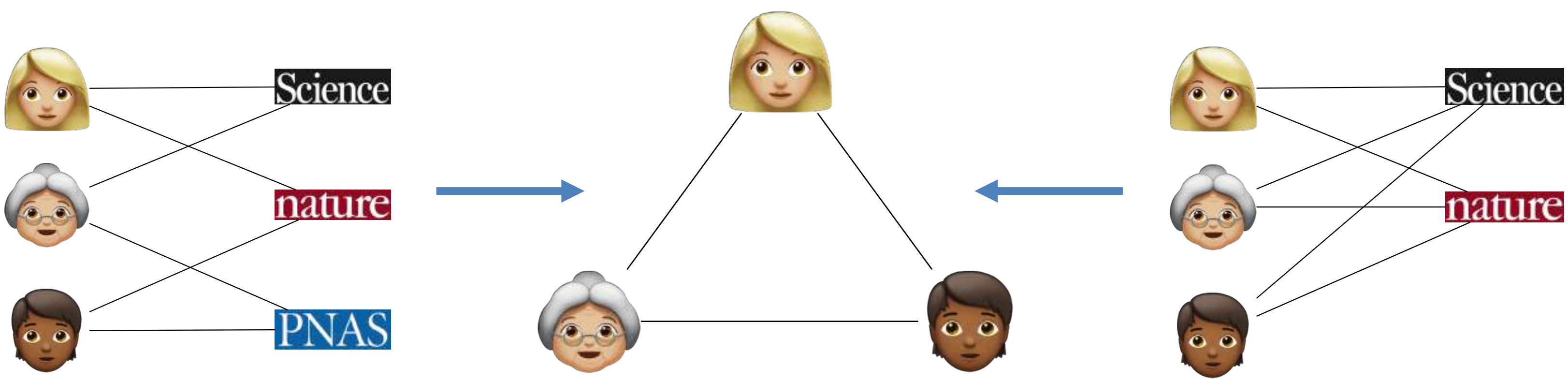


Sariyuce, Seshadhri, Pinar, Catalyurek, TWEB 11(3), 16 (July 2017)

What about other graph types? Bipartite networks (One-to-many relations)?

- Author-paper, word-document, actor-movie…
  - Bipartite in nature, no triangle
- Usually project bipartite to unipartite
  - Author-paper → Co-authorship

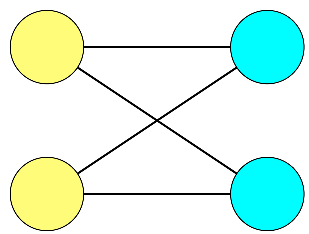
$\lvert E\rvert$	$\lvert E_p\rvert$
58.6K	95.1K
30.7K	84.8K
440.2K	44.5M
96.7K	336.5K
5.6M	157.5M
92.8K	2.0M

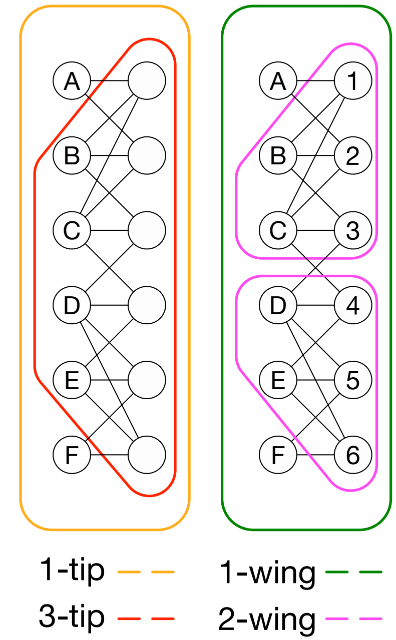


  - |E| explodes! Information lost!
- Find dense regions directly on bipartite graph!

What is the “triangle” in a bipartite network?

- Focus on the smallest non-trivial structure
  - (2, 2)-biclique, or butterfly



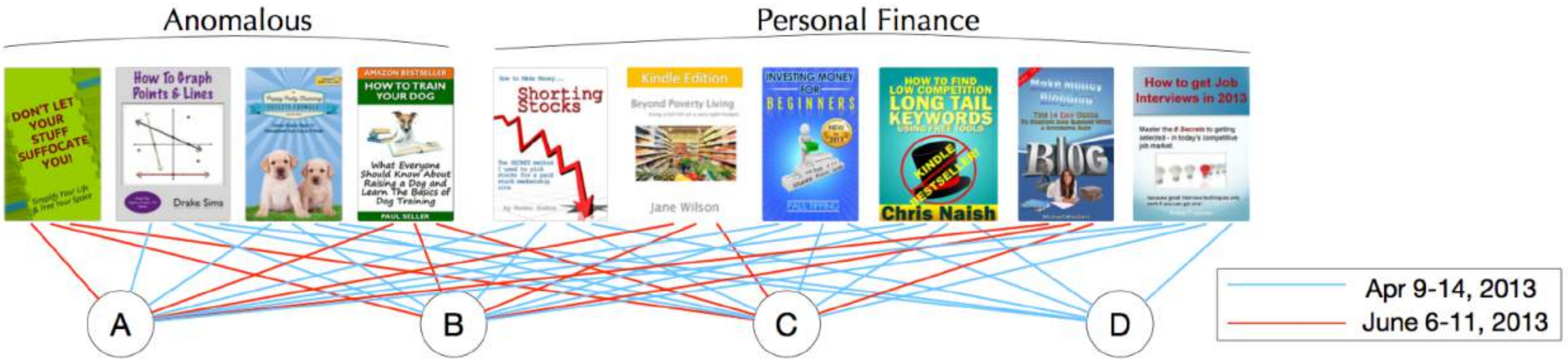


- Vertex-butterfly, edge-butterfly relations
  - *k*-tip: Each vertex has ≥ *k* butterflies
  - *k*-wing: Each edge has ≥ *k* butterflies

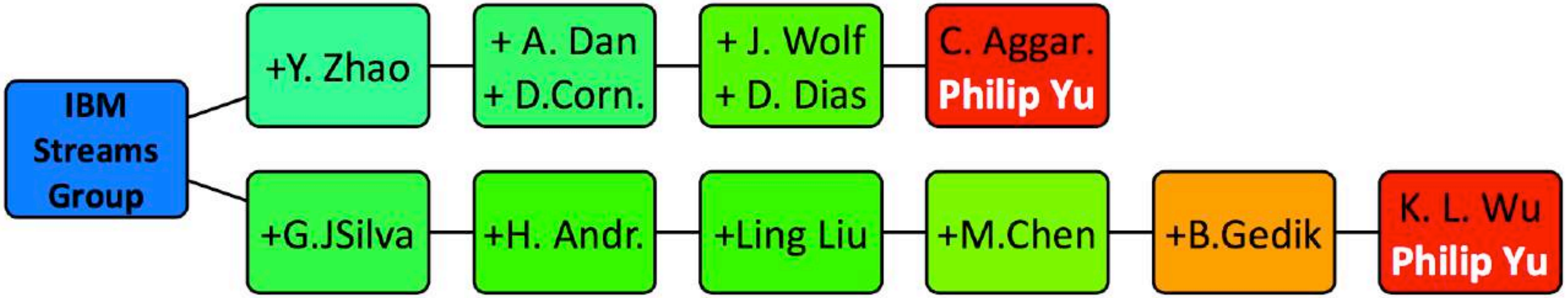
Sariyuce & Pinar, WSDM, 2018

Applications

- Amazon Kindle dataset (users rate books)

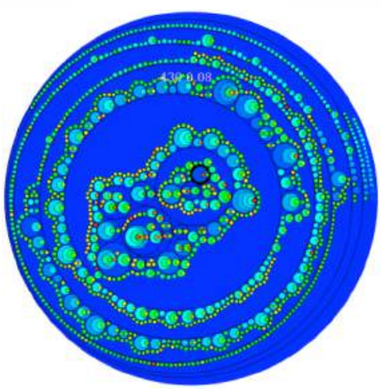


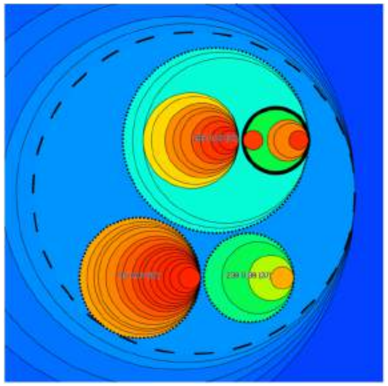
- Author-paper relations at top DB conferences



Conclusion

- Introduced the nucleus decomposition
  - Generalizes *k*-core and *k*-truss; and extend
- Network analysis by the nucleus hierarchy
  - Fast tools
  - Visualization





- Considering bipartite real-world graphs
  - Butterfly is the ‘triangle’ in bipartite graphs

Thanks!

erdem@buffalo.edu

http://sariyuce.com