2018 IEEE International Conference on Data Mining

Summarizing Graphs at Multiple Scales: New Trends

Danai Koutra
University of Michigan

Jilles Vreeken
CISPA Helmholtz Center for Information Security

Francesco Bonchi
ISI Foundation

Roadmap

- 1:30-1:45pm Introduction [Jilles]
- 1:45-2:50pm Network-level Summaries [Francesco]
- 2:55-3:20pm Multi-network Summaries [Danai]
- 3:20-3:40pm ——— *break* ———
- 3:40-4:05pm Multi-network Summaries [Danai]
- 4:10-4:40pm Node-level Summaries [Jilles]
- 4:40-4:50pm Conclusion [Jilles]

Part III:
Local Summarization

Jilles Vreeken

Why do we want a summary?

We want to gain insight in the structure of the data

- capturing the key aspects of the data,
- in easily interpretable terms,
- without redundancy

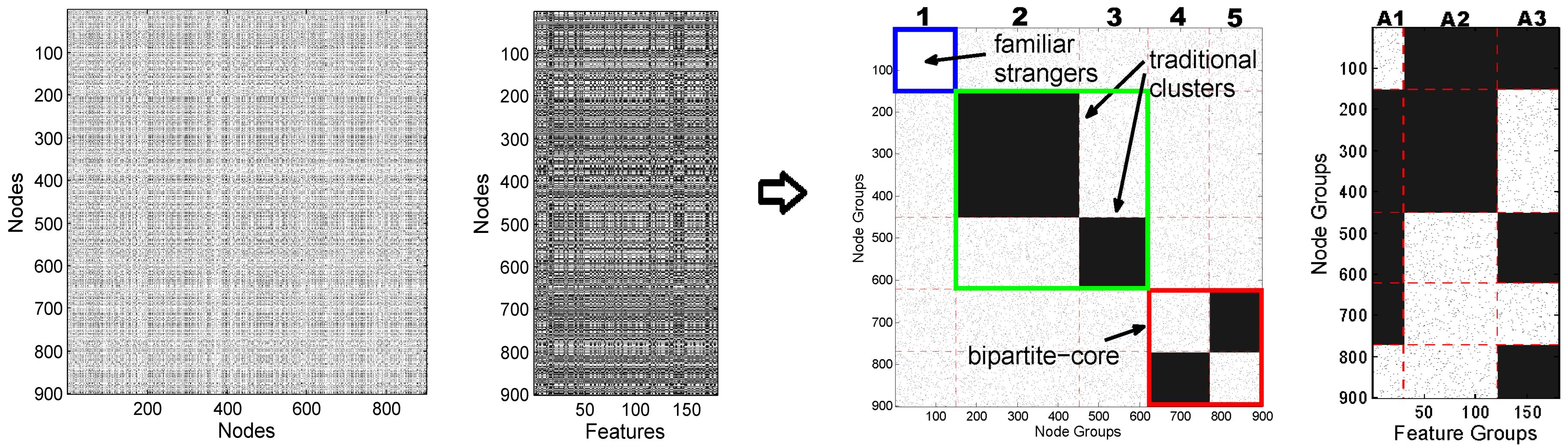
The techniques we've seen so far aim at this

- but, do they really deliver? in all interesting cases?

All deliver one single summary for all of the data

- what do we lose by explaining all the data at once?

Nodes with 'Descriptions'

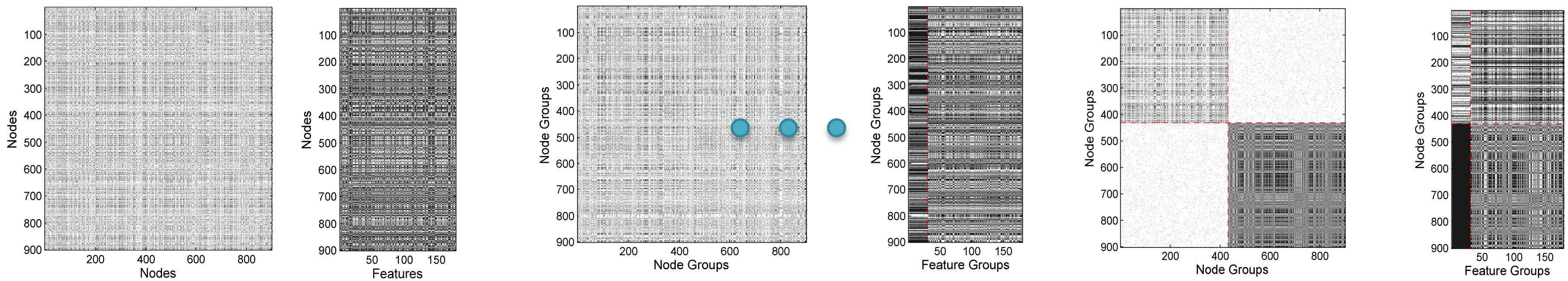


Parameter-free Identification of Cohesive Subgroups in Large Attributed Graphs (Akoglu et al 2012)

- find joint-partition of adjacency matrix and feature matrices
- feature-matrix grid cells can be interpreted as 'descriptions', e.g. 'people with features A1 and A2 but not A3 know each other well', 'people who buy A1, A3, but not A2 all know each other'

[Akoglu et al, '12]

Nodes with ‘Descriptions’



$k = 1, l = 1$ $\quad$ $k = 1, l = 2$ $\quad$ $k = 2, l = 2$

Iteratively add a split on either features or nodes

- after each split, re-arrange nodes and features s.t. sum of entropies over each induced grid cell is minimized
- stop when MDL determines splitting does not provide sufficient gain

[Akoglu et al, ‘12]

Globally not-quite Optimal

Globally summarizing gives an overview, but

- we are not always equally interested in all the data

Moreover, by optimizing a single global objective,

- choices made in how we summarize one part of the data have an effect on other parts of the data
- subgraph G' may be easy to explain, but its locally optimal summary may not fit the global summary well
- globally optimal often means locally suboptimal!

Local Summaries

Why not mine local summaries?

- node groups with exceptional connectivity that come with easy to interpret descriptions

For example,

‘people who watch cat videos often interact’

By not having to care about all the data all the time

- we obtain locally optimal and actionable summaries
- easy to interpret, allowing for alternate explanations

Subgroup Discovery in Graphs

Subgroup Discovery

- given data D and a language $\mathcal{L}$, find those expressions $\sigma \in \mathcal{L}$, such that for a score $s : D \to \mathbb{R}$ we have high $|s(D) - s(\sigma(D))|$

[Klösgen, '96, Friedman, '99]

Subgroup Discovery in Graphs

Subgraph Discovery

- given a graph $G(V, E)$,
  language $\mathcal{L}$ of expressions over nodes (and/or edges),
  e.g. 'nodes with $\boldsymbol{cat\_video = yes}$'

  and a score $\boldsymbol{s}$ over subgraphs,
  e.g. 'average number of edges per node'
- find those expressions $\sigma \in \mathcal{L}$,
  such that the score over induced subgraph G_σ is high,
  e.g. stands out from the score over the whole graph G

Easily understandable, actionable, local summaries

[Atzmueller et al. '16, Pool et al. '14]

Discovering Subgroups in Graphs

Key challenges

- define score s
  - existing measures mainly consider density
    - useful scores are often non-monotone, non-submodular, etc
- define a language $\mathcal{L}$
  - existing languages consider
    - explicit node attributes (cat-video = yes)
    - implicit node attributes (in-degree > 3)
- efficient algorithm to search over $2^{\mathcal{L}}$
  - beam search often used as greedy heuristic without guarantees
  - exact search is possible using branch-and-bound
    if we have an efficiently computable tight optimistic estimator $\bar{s}$

[Knobbe et al. '05, Grosskreutz et al. '08]

Example Subgroup Discovery

Description	#nodes
80s	519
Girl Groups **AND** 80s	215
Atmospheric	171
Synth-Pop	122

Descriptive Communities
found in last.fm

Descriptive Communities
found in Flickr

[Atzmueller et al. 2016, Pool et al. 2014]

References – Local Summaries

Arno Knobbe. **Multi-Relational Data Mining**. ISBN 978-1-58603-661-4, IOS Press, 2005.

Henrik Grosskreutz, Stefan Rüping and Stefan Wrobel. **Tight Optimistic Estimates for Fast Subgroup Discovery**. In Proceedings of the European Conference on Machine Learning and Knowledge Discoveyr in Databases (ECML PKDD), pages 440-456, Springer, 2008.

Martin Atzmueller and Folke Mitzlaff. **Efficient Descriptive Community Mining**. In *Proceedings of 24th International FLAIRS Conference*, AAAI, 2011.

Leman Akoglu, Hanghang Tong, Brendan Meeder, and Christos Faloutsos. **PICS: Parameter-free identification of cohesive subgroups in large attributes graphs.** In *Proceedings of the SIAM Conference on Data Mining* (SDM), pages 439-450, SIAM, 2012.

Simon Pool, Francesco Bonchi, and Matthijs van Leeuwen. **Description-Driven Community Detection**. In *ACM Trans. Intell. Syst. Technol.* 5(2), 28:1-28, ACM, 2014.

Martin Atzmueller, Stephan Dörfel, and Folke Mitzlaff. **Description-oriented community detection using exhaustive subgroup discovery**. In *Info. Sci.* 329, pages 965-984, 2016.

Explain me this…

We do not always care about summarize graph G entirely

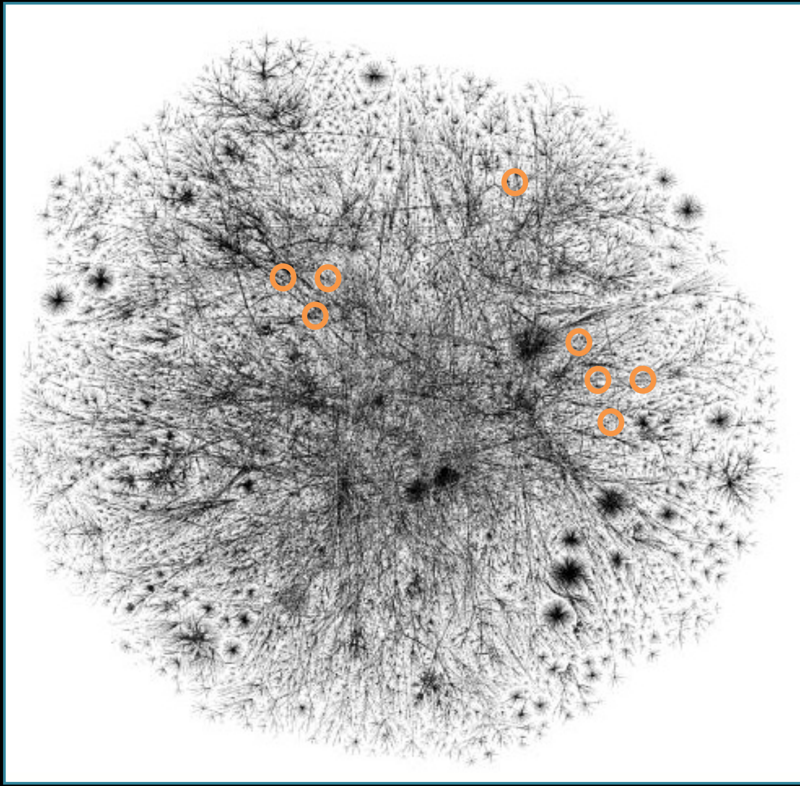
- how to explain nodes $S \subseteq V$ marked by external process?

What can G explain about S ?

- are S close by each other?
- are S segregated?
- how many groups do they form?

How can we connect S using G?

- with "simple" paths
- using "good" connectors

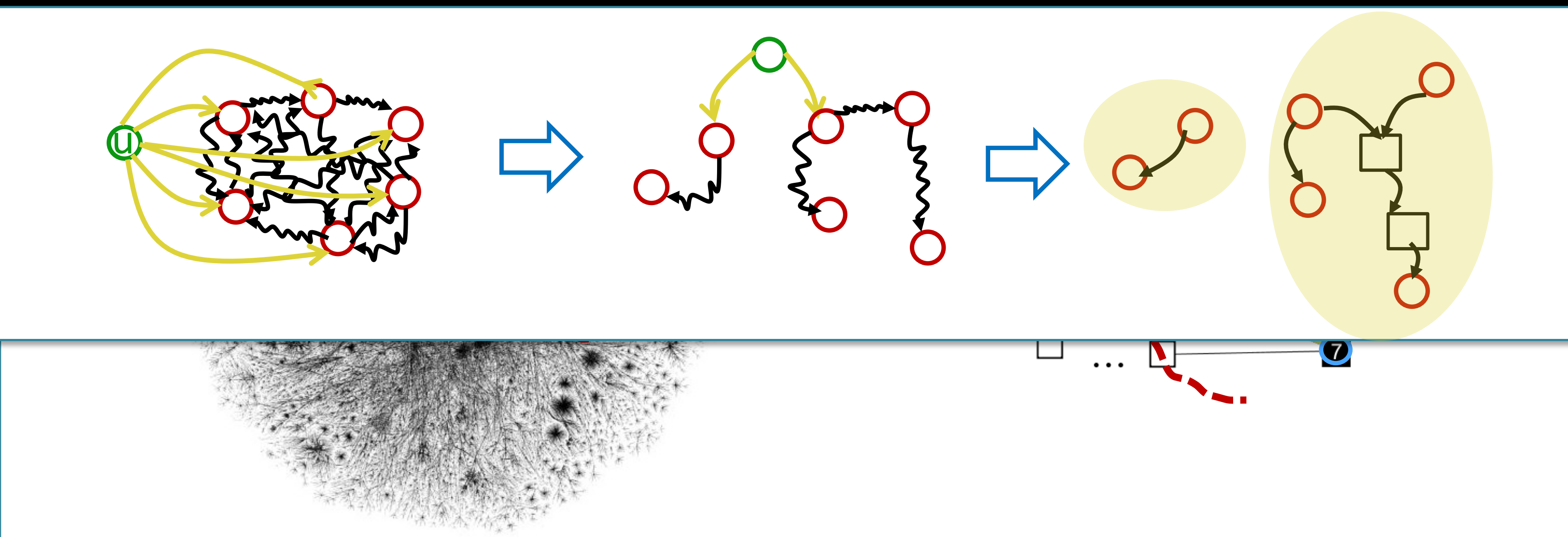


[Akoglu et al. 2013]

Simple Connection Pathways

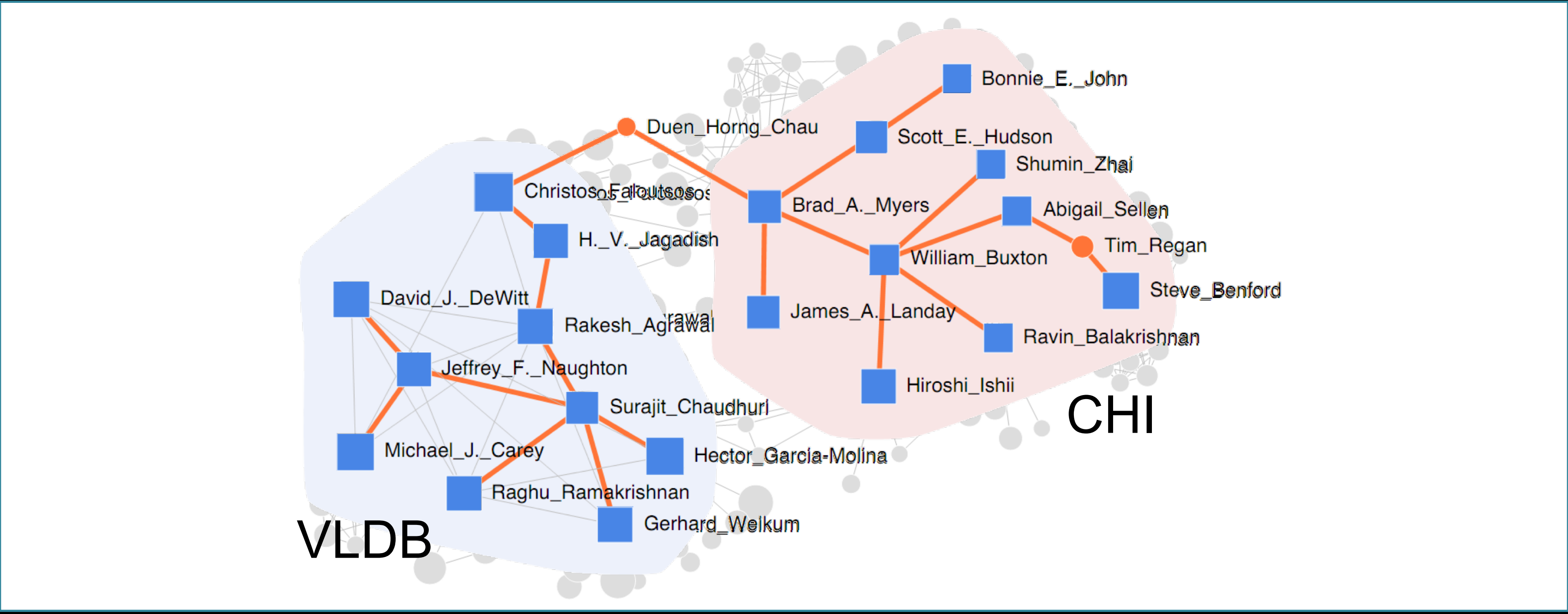
Main idea: use the network structure to explain ***S***

- partition ***S*** into groups of nodes, such that:
  - ✧ "simple" paths connect the nodes in each group, nodes in different groups are "not easily reachable"



[Akoglu et al. 2013]

Example in Co-Authorship Graph



[Akoglu et al. 2013]

Subjectively Interesting

For dense graphs, there are no ‘simple’ paths

- plain path simplicity by MDL does not work (well)

However, some paths are more expected

- for example, paths between recently active nodes
- we can express $\Pr(path)$ using such external information

And mine most informative SteinerTree incrementally

- iteratively add $edge$ with highest $\dfrac{InfContent(edge)}{DescLength(edge)}$

[Adriaens et al. 2017]

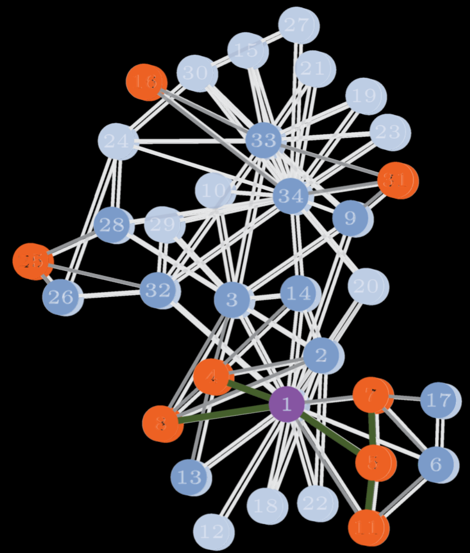
Bump Hunting in the Dark

Why explain all query nodes $S \subseteq V$?

- why not as many as possible with one connected subgraph?

Find connected subgraph G'
with many nodes in S and
few nodes not in S

- i.e. $G' \subseteq G$ is connected, with high $|S \cap V'|$, and low $|V' \setminus S|$
- this is known as discrepancy maximization: NP-hard on graphs
- note the relation to subgroup discovery!



[Gionis et al. 2017]

Bump Hunting in the Dark

Find connected G' with high $|S \cap V'|$ and low $|V' \setminus S|$

- NP-hard, no known approximation algorithms

If graph G is a tree it's easy, but if it's a graph it's hard

- main idea find a tree $G_T \subseteq G$, then find G' on G_T
- linear time heuristics to find G_T based on spanning trees
- variants for full graph access, and for local expansion

Key open questions

- weighing scheme, expansion strategy, stopping criteria
- and, how to expand to other, more refined measures

[Gionis et al. 2017]

Minimally Inefficient

Connectedness of G' restricts usefulness

- instead, find that set of nodes $C \subseteq V \setminus S$ such that induced subgraph $G' = G[S \cup C]$ is cohesive

Cohesiveness relates to reachability

- if G' is not connected, shortest path may be infinite
- efficiency of a graph defined as

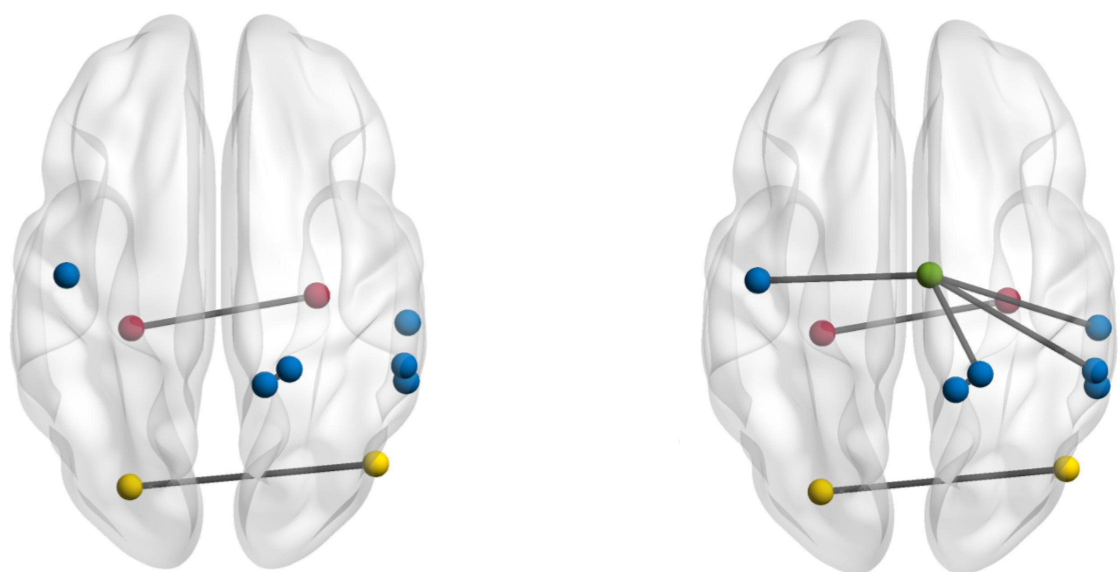
$$\mathcal{E}(G) = \frac{1}{|V|(|V|-1)} \sum_{\substack{u,v \in V \\ u \neq v}} \frac{1}{d_G(u,v)}$$

[Ruchansky et al. 2017]

Minimally Inefficient

Minimium Inefficiency Subgraph problem

- find those nodes $C \subseteq V \setminus S$ such that induced subgraph $G' = G[S \cup C]$ is minimally inefficient
- NP-hard, not known to be approximable: greedy heuristic



[Ruchansky et al. 2017]

References – Connected Nodes

Leman Akoglu, Jilles Vreeken, Hanghang Tong, Nikolaj Tatti, and Christos Faloutsos. **Mining Connection Pathways for Marked Nodes in Large Graphs**. In *Proceedings of the 13th SIAM International Conference on Data Mining* (SDM), pages 37-45, SIAM, 2013.

Stephan Seufert, Klaus Berberich, Srikanta J. Bedathur, Sarath Kumar Kondreddi, Patrick Ernst, and Gerhard Weikum. **ESPRESSO: Explaining Relationships between Entity Sets**. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management* (CIKM), pages 1311-1320, ACM, 2016.

Aristides Gionis, Michael Mathioudakis, and Antti Ukkonen. **Bump hunting in the dark: Local discrepancy maximization on graphs**. *IEEE Trans. Knowl. Data Eng.*, 29(3):529–542, IEEE, 2017.

Florian Adriaens, Jefrey Lijffijt, and Tijl De Bie. **Subjectively interesting connecting trees**. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases* (ECML PKDD), pages 53–69, Springer, 2017.

Natali Ruchansky, Francesco Bonchi, David Garcia-Soriano, Francesco Gullo, and Nicolas Kourtellis. **To Be Connected, or Not to Be Connected: That is the Minimum Inefficiency Subgraph Problem**. In *Proceedings of the 26th ACM Conference on Information and Knowledge Management* (CIKM), pages 879-888, ACM, 2017.

A Picture Says More Than…

Ideally a summary is easy to understand

- most solutions provide pretty bad presentation
- and, no support for exploration of the summary space

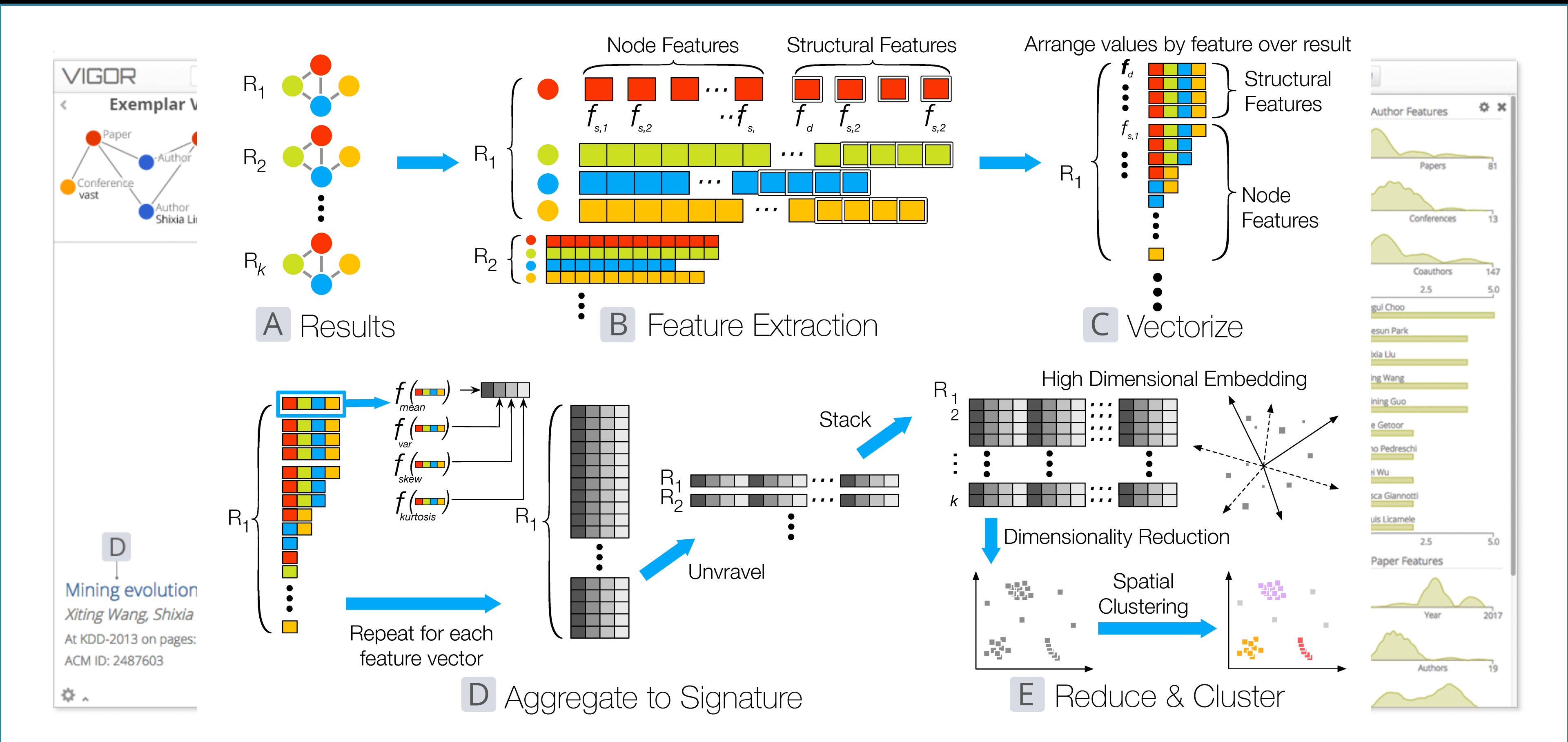
Key difficulties

- how should we present a summary?
- how can we interact with it?

Only few visual summary exploration tools exist

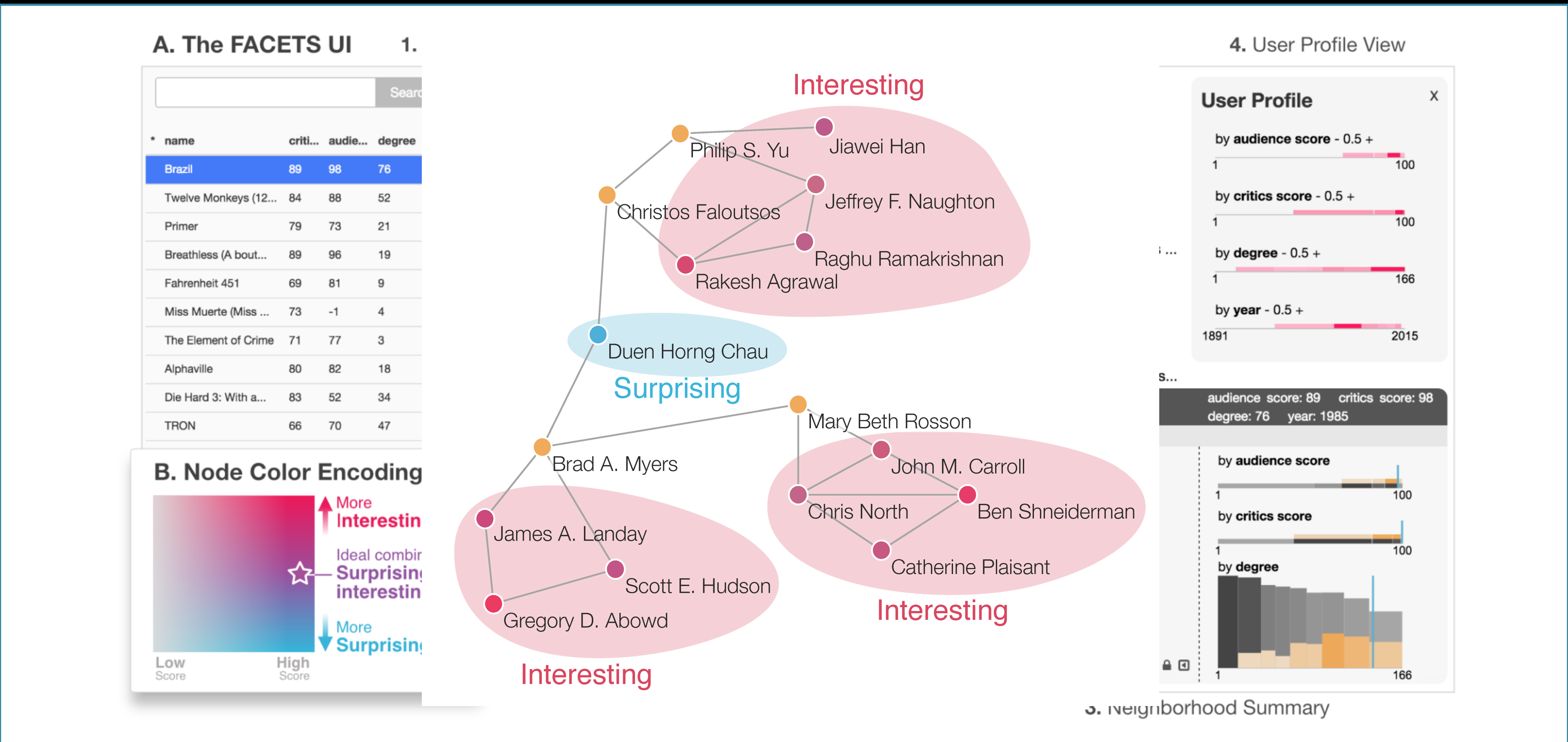
- and out of those, we only cover two

Interactive Exploration of Graph Query Results



[Pienta et al. 2018]

Graph Exploration



[Pienta et al. 2017]

References – Visual Exploration

Bryan Perozzi and Leman Akoglu. **Discovering communities and anomalies in attributed graphs: Interactive visual exploration and summarization**. In *ACM Trans. Knowl. Disc. Data Min*, 12(2):24:1–24:40, ACM, 2018.

Robert Pienta, Fred Hohman, Alex Endert, Acar Tamersoy, Kevin A. Roundy, Christopher S. Gates, Shamkant B. Navathe, and Duen Horng Chau. **VIGOR: interactive visual exploration of graph query results**. In *IEEE Trans. Vis. Comput. Graph*., 24(1):215–225, IEEE, 2018.

Robert Pienta, Minsuk Kahng, Zhiyuan Lin, Jilles Vreeken, Partha P. Talukdar, James Abello, Ganesh Parameswaran, and Duen Horng Chau. **FACETS: adaptive local exploration of large graphs**. In *Proceedings of the SIAM International Conference on Data Mining* (SDM), pages 597–605, 2017.

Boxin Du, Si Zhang, Nan Cao, and Hanghang Tong. 2017. **FIRST: Fast Interactive Attributed Subgraph Matching**. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), pages 1447-1456, ACM, 2017.

Roadmap

1:30-1:45pm	Introduction	[Jilles]
1:45-2:50pm	Network-level Summaries	[Francesco]
2:55-3:20pm	Multi-network Summaries	[Danai]
3:20-3:40pm	*break*	
3:40-4:05pm	Multi-network Summaries	[Danai]
4:10-4:40pm	Node-level Summaries	[Jilles]
4:40-4:50pm	**Conclusion**	[Jilles]

What Have We Seen

Summarization of Single Networks

- a lot of work done, in many different angles

Summarization of Multiple Networks

- not so much work done, big open problems

Summarization of Sets of Nodes

- very little work done, very interesting challenges

Single Networks

Single network summarization is challenging

- *how to decide what is important? Lossy or lossless? What is the goal of the summary? How to keep things tractable?*

Main focus: unattributed undirected networks

- simple problem is already hard enough, covers many settings

There exists, but only very limited work on

- attributed, directed, or signed networks

Big challenges, huge opportunities!

Multiple Networks

- Multi-network summarization is more challenging than network-level summarization
  - ✧ *How to reduce re-computations? pick the right temporal granularity? handle node additions / deletions? make methods scale to multiple networks?*
- Main focus: temporal networks
  - ✧ Applying static methods on snapshots is not sufficient
  - ✧ Different models: static snapshots / tensor, graph stream
- Very limited work on
  - ✧ attributed temporal networks
  - ✧ multiple disparate networks
- "One size does not fit all"! We should be thinking about tailored summaries: domain-specific, personalized, query-driven etc.

Big challenges, huge opportunities!

Nodes

Taking a local rather than a global perspective

- descriptions of subgraphs, much easier to understand
- no global choices that affect locally optimality!

Surprisingly little work done

- discovering explainable subgraphs
- explaining node sets
- interactive exploration and interaction

Big challenges, huge opportunities!

Open Research Problems

- Richer data (attributed, temporal, spatio-temporal, multilayer)
- Real-time graph summarization (streaming, incremental)
- Summary maintenance
- Evaluation (which metrics?)
- Automated insight extraction (explanation, storytelling)

Open Research Problems

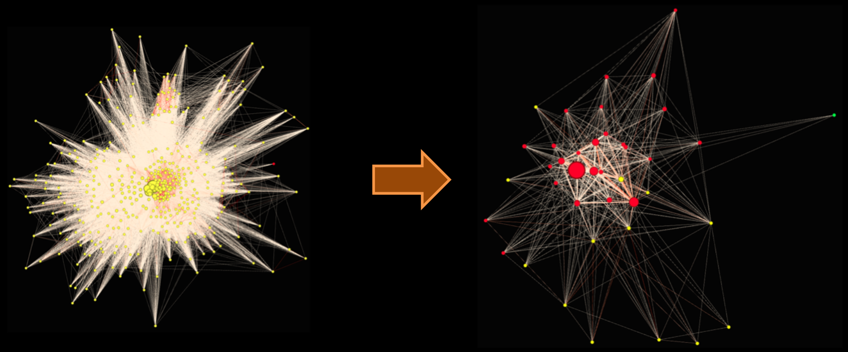
- Scalable, high quality attribute-aware summaries
- Application-driven (domain-dependent) summarization
- Summarization of uncertain graphs
- Summarizing a set of graphs
- Differential summaries on massive networks

Conclusions

Graph summarization is important and has impact:

- reduction of data volume + storage
- speedup of algorithms + queries
- interactive analysis
- noise elimination (patterns)

There is a lot of potential for
high-impact contributions!

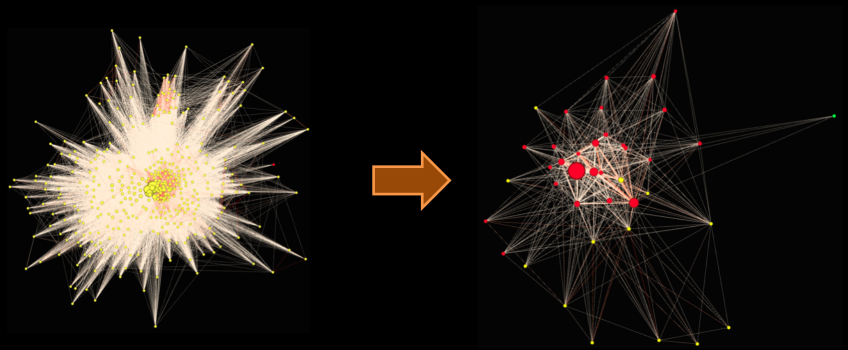


Conclusions

Graph summarization

- no unique approach or notion
- no widely accepted objective function
- no standard evaluation measure or benchmark
- highly domain and application dependent

There is a lot of potential for
high-impact contributions!



Resources

D. Koutra "Summarizing Large-Scale Graph Data: Algorithms, Applications & Open Challenges" (SDM 2017 - tutorial)

A. Khan, S.S. Bhowmick, F. Bonchi "Summarizing Static and Dynamic Big Graphs" (VLDB 2017 - tutorial)

Liu, Safavi, Dighe, Koutra "Graph Summarization Methods and Applications: A Survey" (ACM Comp. Surv. 2018)