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

About the presenters

Danai
Koutra
University of Michigan

Jilles
Vreeken
CISPA Helmholtz Center
for Information Security

Francesco
Bonchi
ISI Foundation

About this tutorial

- ~3 hours
- Extensive but incomplete overview of related works
  - naturally (quite) a bit biased
- Partially based on:

  Liu, Safavi, Dighe, Koutra.
  Graph Summarization Methods and Applications: A Survey. ACM Comp. Surv. 51, 3, Article 62. ACM, 2018.
  https://dl.acm.org/citation.cfm?id=3186727

Graph Summarization Methods and Applications: A Survey

YIKE LIU, TARA SAFAVI, ABHILASH DIGHE, and DANAI KOUTRA, University of Michigan, Ann Arbor

While advances in computing resources have made processing enormous amounts of data possible, human ability to identify patterns in such data has not scaled accordingly. Efficient computational methods for condensing and simplifying data are thus becoming vital for extracting actionable insights. In particular, while data summarization techniques have been studied extensively, only recently has summarizing interconnected data, or *graphs*, become popular. This survey is a structured, comprehensive overview of the state-of-the-art methods for summarizing graph data. We first broach the motivation behind and the challenges of graph summarization. We then categorize summarization approaches by the type of graphs taken as input and further organize each category by core methodology. Finally, we discuss applications of summarization on real-world graphs and conclude by describing some open problems in the field.

62

CCS Concepts: • **Mathematics of computing** → **Graph algorithms**; • **Information systems** → **Data mining**; **Summarization**; • **Human-centered computing** → *Social network analysis*; • **Theory of computation** → *Unsupervised learning and clustering*; • **Computing methodologies** → *Network science*;

Additional Key Words and Phrases: Graph mining, graph summarization

ACM Reference format:
Yike Liu, Tara Safavi, Abhilash Dighe, and Danai Koutra. 2018. Graph Summarization Methods and Applications: A Survey. *ACM Comput. Surv.* 51, 3, Article 62 (June 2018), 34 pages.
https://doi.org/10.1145/3186727

1 INTRODUCTION

As technology advances, the amount of data that we generate and our ability to collect and archive such data both increase continuously. Daily activities like social media interaction, web browsing, product and service purchases, itineraries, and wellness sensors generate large amounts of data, the analysis of which can immediately impact our lives. This abundance of generated data and its velocity call for data summarization, one of the main data mining tasks.

Since summarization facilitates the identification of structure and meaning in data, the data mining community has taken a strong interest in the task. Methods for a variety of data types

Y. Liu and T. Safavi contributed equally to this article.

This material was based on work supported in part by the National Science Foundation under grant IIS 1743088, Trove, and the University of Michigan. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or other funding parties. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

Authors' addresses: Y. Liu, T. Safavi, A. Dighe, and D. Koutra, Bob and Betty Beyster Building, 2260 Hayward St, Ann Arbor, MI 48109; emails: {yikeliu, tsafavi, adighe, dkoutra}@umich.edu.

https://doi.org/10.1145/3186727




What we won't cover

For example, we will not discuss summarizing

- itemsets and association rules [Liu et al. 1999; Mampaey et al. 2011a,b; Ordonez et al. 2006; Wang and Parthasarathy 2006; Yan et al. 2005]
- event sequences [Garriga 2005; Kiernan et al. 2009; Tatti et al. 2012],
- spatial data [Lin et al. 2003],
- transactions and multi-modal databases [Chandola & Kumar 2005; Cordeiro et al. 2010; Shneiderman 2008; Wang et al. 2004; Xiang et al. 2010],
- data streams and time series [Cormode et al. 2005; Palpanas et al. 2008],
- video and surveillance data [Damnjanovic et al. 2008; Pan et al. 2004]

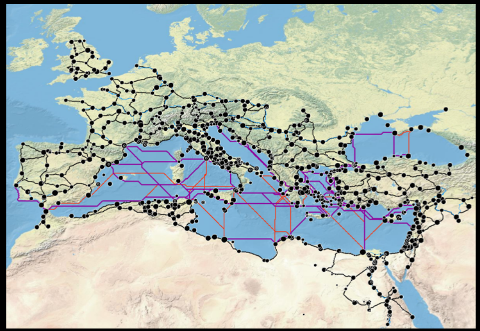
Schedule

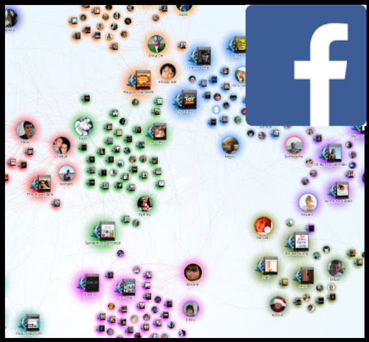
- 1:30-1:45pm Introduction [Jilles]
- 1:45-2:50pm Network-level Summaries [Francesco]
- 2:55-3:20pm Multi-network Summaries [Danai]
- 3:20-3:40pm ——— *break* ———
- 3:40-4:05pm Multi-network Summaries [Danai]
- 4:10-4:40pm Node-level Summaries [Jilles]
- 4:40-4:50pm Conclusion [Jilles]

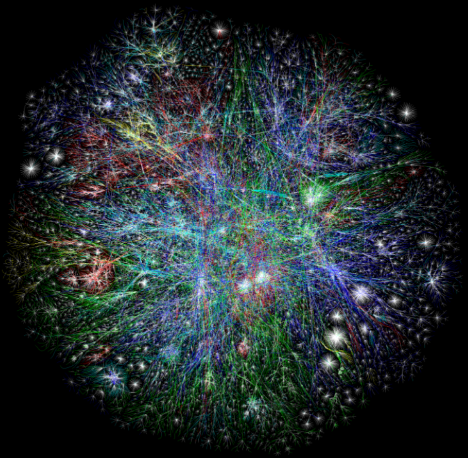
Roadmap

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]

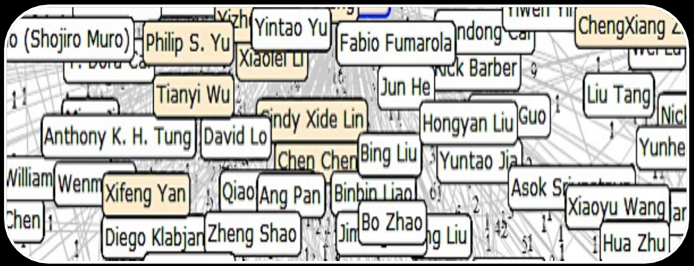
Graph Data

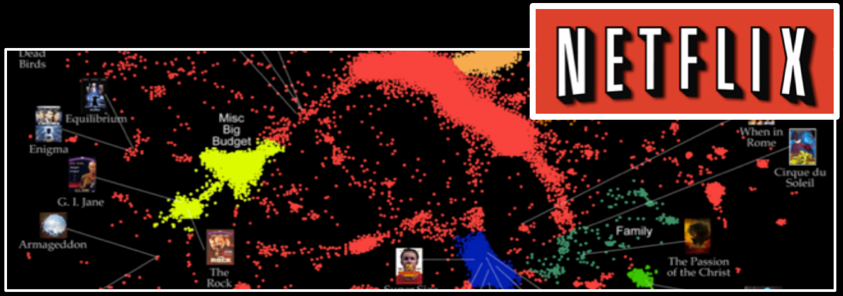




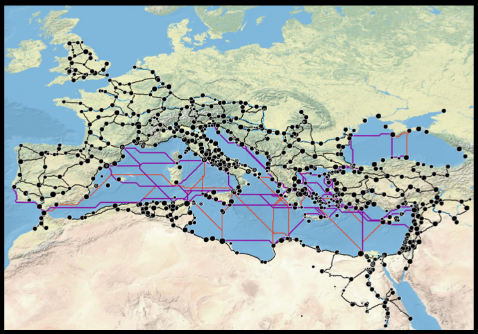


dblp

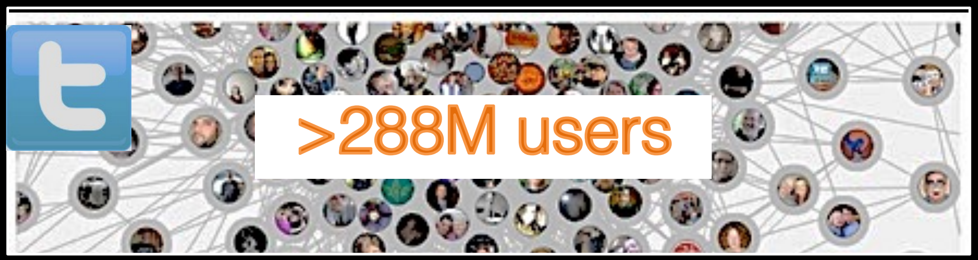




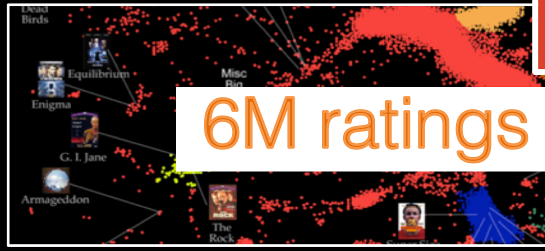
LARGE-scale Graph Data



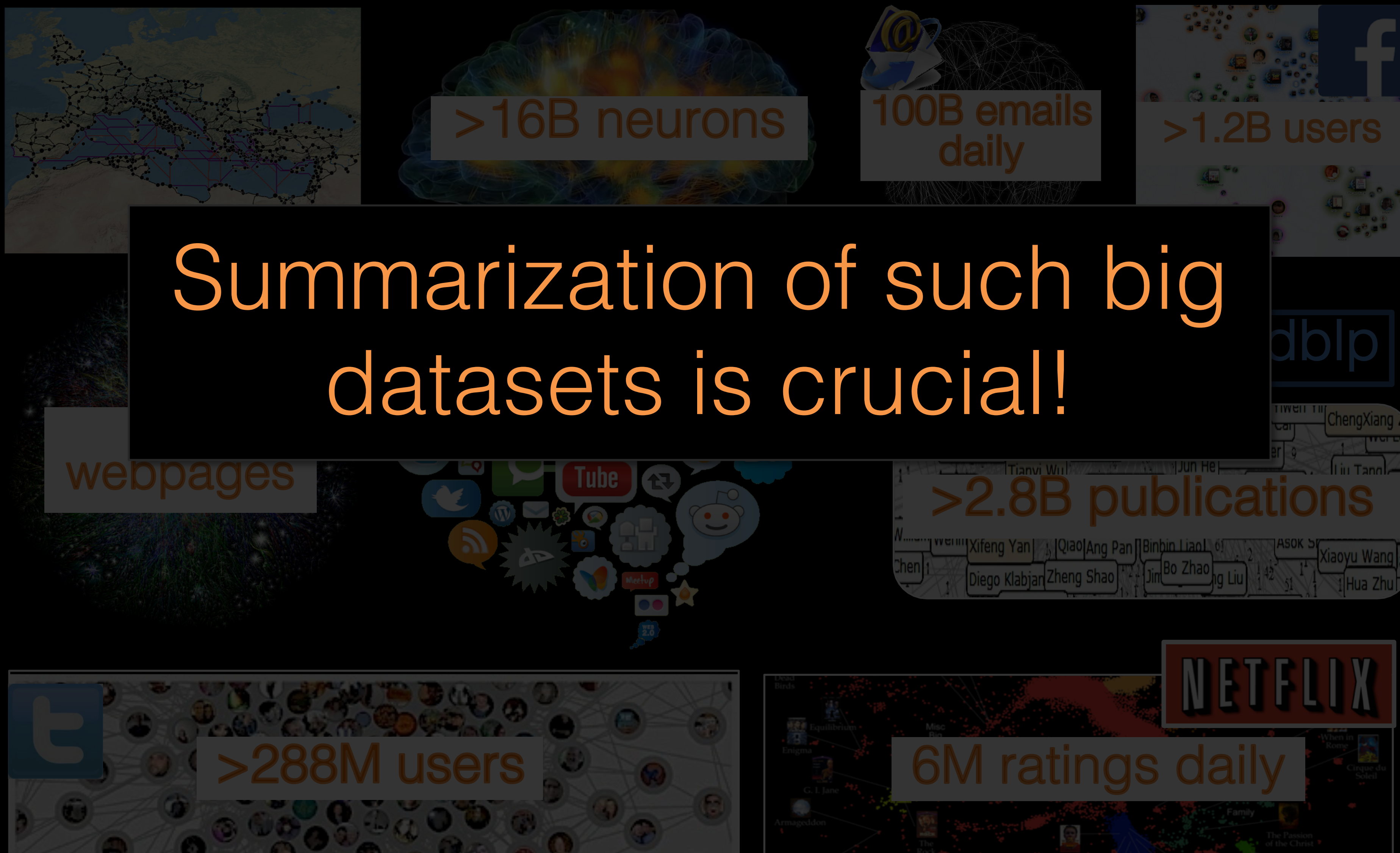
dblp



NETFLIX



LARGE-scale Graph Data



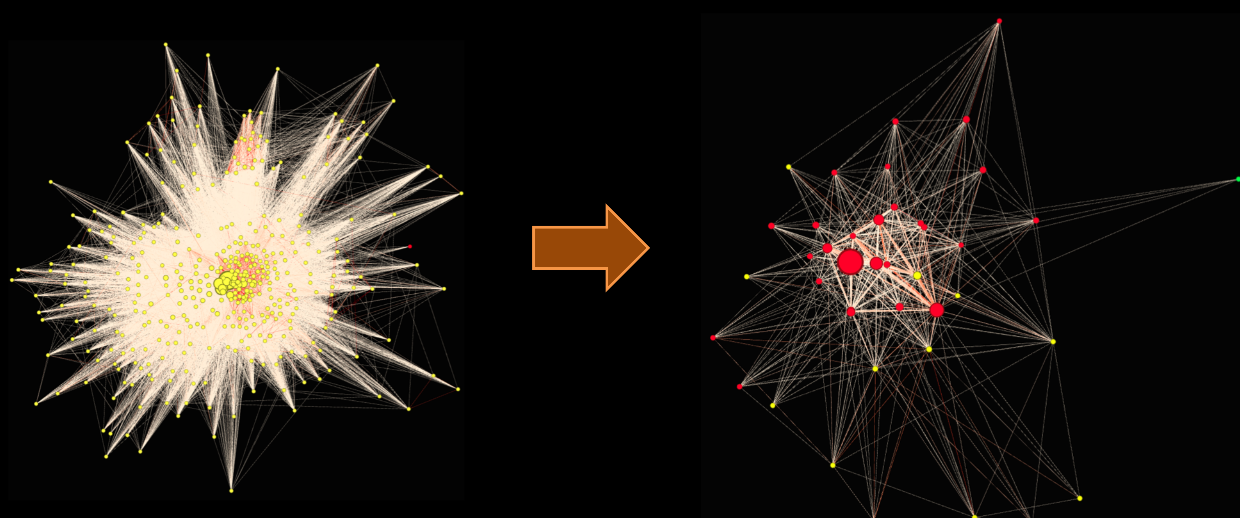
Summarization of such big datasets is crucial!

What is graph summarization?

(or coarsening or aggregation)

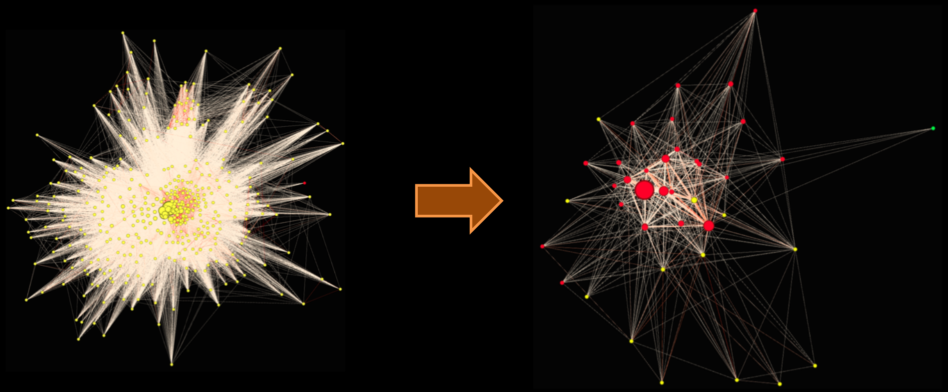
It seeks to find

- a short representation of the input graph,
- often in the form of a summary or sparsified graph,
  - ✧ which reveals patterns in the original data and preserves specific structural or other properties, depending on the application domain.



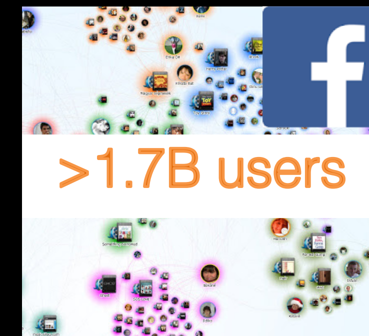
Why graph summarization?

- Reduction of data volume + storage
  - ✧ e.g., fewer I/O operations
- Speedup of algorithms + queries
- Interactive analysis
- Noise elimination
  - ✧ reveals patterns

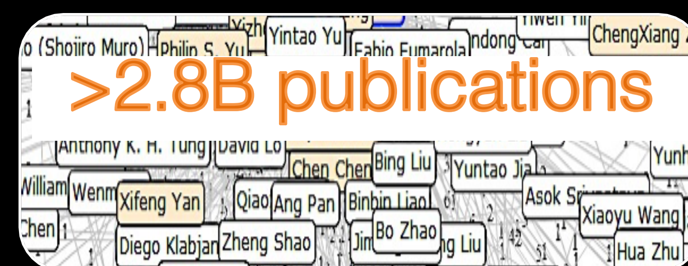


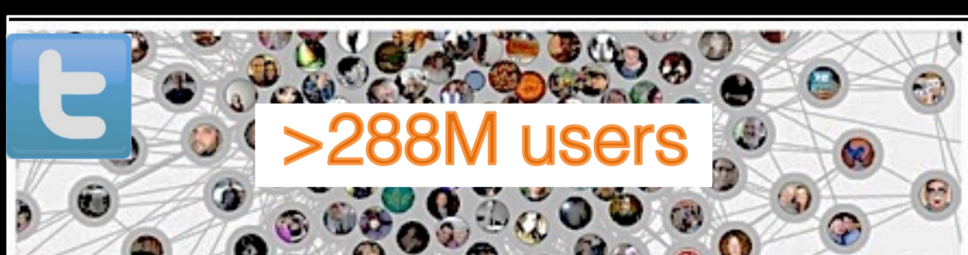
Challenges

- Volume of data



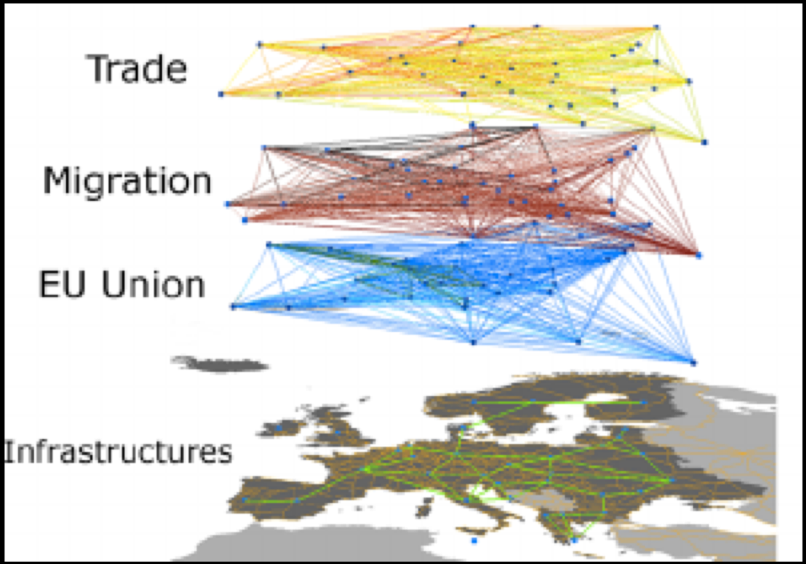
dblp





Challenges

- Volume of data
- Complexity of data
  - dependencies, side information (attributes, …)



Challenges

- Volume of data
- Complexity of data
- Definition of interestingness / importance
  - subjective, application-dependent

Challenges

- Volume of data
- Complexity of data
- Definition of interestingness / importance
- Changes over time

Challenges

- Volume of data
- Complexity of data
- Definition of interestingness / importance
- Changes over time
- Evaluation
  - what makes a summary a good summary?

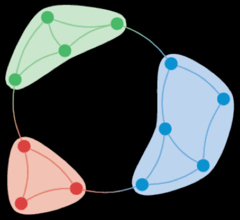
How to evaluate a summary?

There exists no universal summarization metric

- Compression-based:
  - ✧ minimize number of bits without losing much information, reduce # nodes / edges

- Query-oriented (e.g., reachability):
  - ✧ accuracy vs. runtime

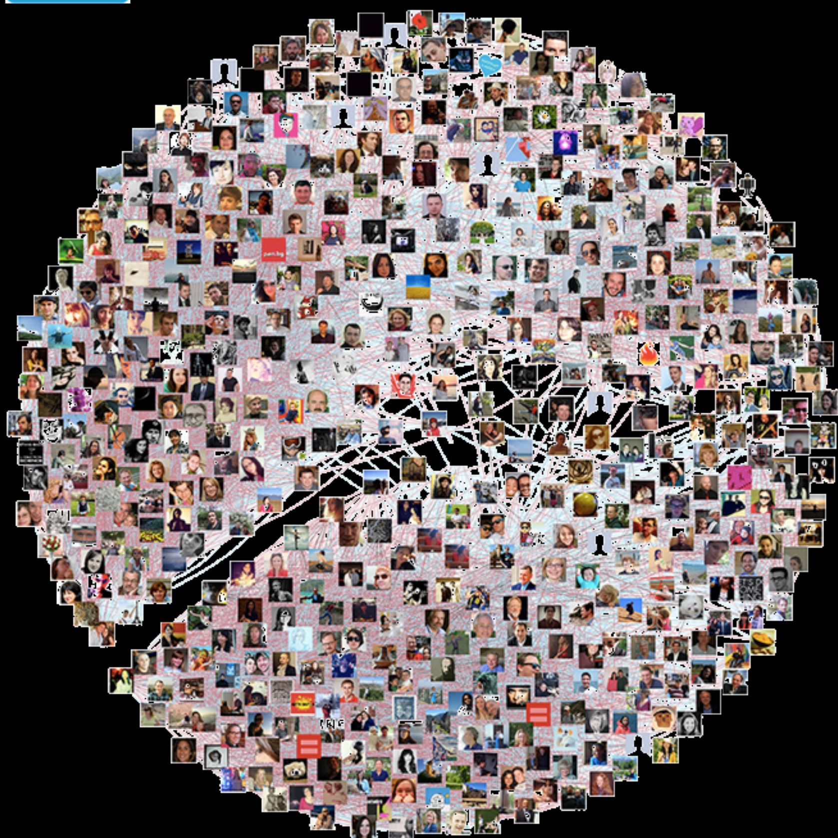
- Clustering-oriented:
  - ✧ maintain community structure



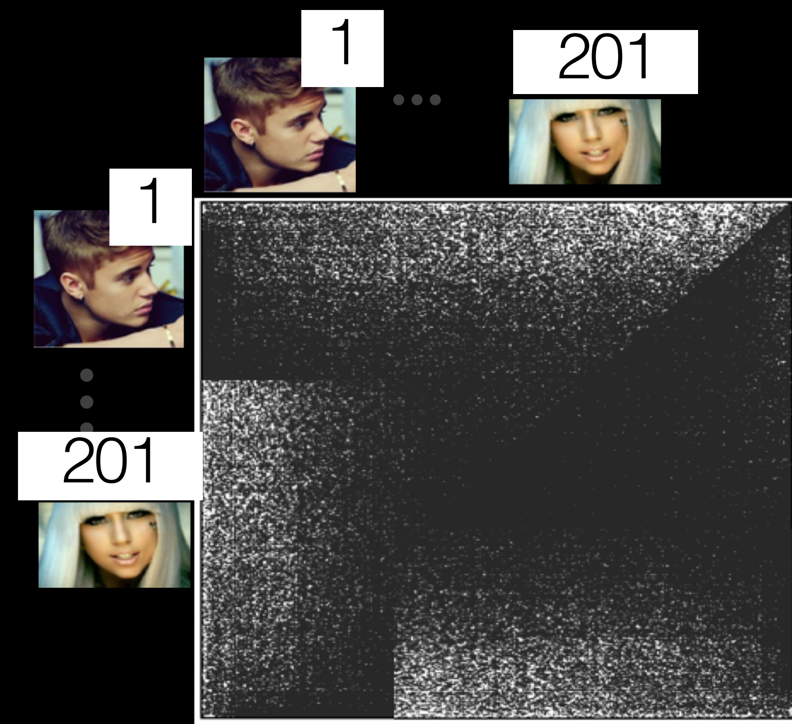
- Quality-based measures:
  - ✧ “interestingness”, reconstruction error

Graph Representation

Graph G_A



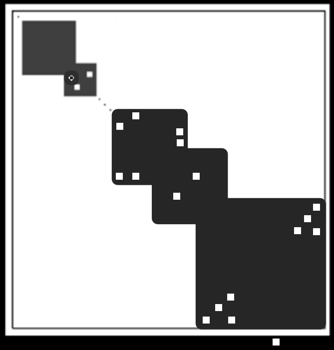
Adjacency matrix A



Types of graphs

- Weighted / Unweighted
  - (w) # of msg
  - (w) # of phonecalls
  - (w) distance
  - (u) friendship
- Directed / Undirected
  - (d) Caller, callee
  - (d) Who-follows-whom
  - (u) Friendship (FB)
- Labeled / Unlabeled
- Homogeneous / Heterogeneous

Undirected graph
(symmetric)



Challenges

- Volume of data
- Complexity of data
- Definition of interestingness / importance
- Changes over time
- Evaluation
- What should be summarized?
  - ✧ we’re not always interested in the whole graph,
  - ✧ globally optimal may mean locally suboptimal

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]