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:
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]
- 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
LARGE-scale Graph Data

>16B neurons

100B emails daily

>1.7B users

50B webpages

>2.8B publications

>288M users

6M ratings daily
LARGE-scale Graph Data

Summarization of such big datasets is crucial!

- >16B neurons
- >1.2B users
- 100B emails daily
- >2.8B publications
- >288M users
- 6M ratings daily
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
  - 100B emails daily
  - >1.7B users
  - >2.8B publications
  - 50B webpages
  - >16B neurons
  - >288M users
  - 6M ratings daily
Challenges

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

![Diagram showing trade, migration, EU Union, and infrastructures in Europe.](image)
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
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]