Outline

- Network abstraction
- Network abstraction applications in Bioinformatics
- Future plans
Network abstraction

- A network (or graph) is an abstract representation of linked data: nodes represent objects, and edges represent connections between objects.
Network abstraction

- **Network abstraction** transforms a large graph into a smaller and simpler graph with respect to a possible application.

- We proposed network abstraction as a research area.
Network abstraction includes various types of existing approaches in data mining, and also raises new research problems.

Related work:

- Graph sampling [Rafiei 2005]
- Relevant subgraph extraction [Faloutsos et al. 2004, Hintsanen et al. 2008; 2010]
- Graph partitioning [Girvan and Newman 2002]
- Graph summarization [Navlakha et al. 2008, Tian et al. 2008]
- Etc.
Three main research questions

- Q1: What kind of operations and methods can we use to abstract a graph?

- Q2: How can we measure the quality of an abstracted graph?

- Q3: How useful is an abstracted graph?
What kind of operations can we use to abstract a graph?

We developed a taxonomy for network abstraction approaches.

- Three types of operations:
  - Prune
  - Partition
  - Replacement

- Two viewpoints:
  - Objective methods
  - Subjective methods
Prune:

Partition:

Replace:
Research topics

- **Network simplification**
  - Lossless network simplification
  - Lossy network simplification

- **Network compression**
  - Weighted graph compression
Lossless network simplification

- Removal of edges that are *redundant* to the whole network.
- Definition: An edge is redundant iff its removal does not affect the best path quality for any pair of nodes.
- How can we measure the quality of an abstracted network?

\[ Q(V, E) = Q(V, E \setminus E') \]
Lossless network simplification

Original graph

Result graph
Lossy network simplification

- Removal of edges that are *less important* to the whole graph.
- Given a graph \( G=(V, E) \) and a number of edges to be removed.
- Produce a simplified graph \( H=(V, E \setminus E') \) that *best* maintains the overall graph connectivity.
- Maximize: \( Q(V, E \setminus E') / Q(V, E) \).
Lossy network simplification

Original graph

Result graph
Research topics

- Network simplification
  - Lossless network simplification
  - Lossy network simplification

- Network compression
  - Weighted graph compression
Weighted Graph Compression

- Grouping nodes that connect with similar neighbors.

Original graph

Result graph
Weighted Graph Compression

- Compression with (nodes and edges) deletion
Weighted Graph Compression

- Given a weighted $G$, a compression ratio.

- Produce a compressed graph $S$ that contains approximately the same information as the graph $G$. 
Weighted Graph Compression

- Optimal superedge weight

\[ w'(\{u', v'\}) = \frac{\sum_{\{u,v\} \in u' \times v'} w(\{u, v\})}{|u'| \cdot |v'|} \]

- Bounds for distances between graphs

\[ \text{dist}(G_1, G_2) \leq \sqrt{\alpha} \sum_{\{u,v\} \in V \times V} d_{\text{max}}(G_1, G_2)^2 \]

- A bound on distances between nodes

\[ d_{\text{max}}(u, v; S) = \max\{ \max_{x: \{u, x\} \text{ or } \{v, x\} \in E} (|Q(\lambda(u, x; S) - Q(\lambda(v, x; S)|)) \]

\[ |Q(\lambda(u, u; S) - Q(\lambda(v, v; S)|), \]

\[ |Q(\lambda(u, u; S) - Q(\lambda(u, v; S)|), \]

\[ |Q(\lambda(v, v; S) - Q(\lambda(u, v; S)| \}. \]
Conclusions

- We defined and structured the field of network abstraction.

- We proposed two computational problems, and developed novel methods and tools.

  - Papers were published in data mining top conferences: KDD, ICDM, IDA.
Bioinformatics

The Use of Weighted Graphs for Large-Scale Genome Analysis
Metabolism
Data

- Super-metabolic graph
  - 192 pathways
  - 2,365 enzymes, and 43,627 metabolite edges

- Species with complete genomes
  - 108 Archaea species
  - 1,287 Eubacteria species

- Pair-wise comparison: 972,315 times!
The use of weighted graph

- **Idea**: integrate multiple genomes into one graph through the use of weighted graphs.
Three types of graphs with weights on enzymes:

- Taxonomic weights - summarize phylogenetic importance.
- Isoenzymatic weights - summarize enzymatic variety/redundancy.
- Sequence-similarity weights - summarize sequence conservation.
We applied our weighted graph compression method to metabolic networks.

The results provided the evidences for the limits to the contingency of evolution, and demonstrated the utility of our graph compression approach.

This work is published in PLOS One.
Future plans
Further work in network analysis

Applications in Multimedia
  * Personalized trip recommendation
Acknowledgement

- Prof. Hannu Toivonen  (University of Helsinki, Finland)
- Prof. Ross D. King  (University of Manchester, UK)
- Prof. Guoping Qiu  (University of Nottingham-Ningbo, China)
- Aleksi Hartikainen  (University of Helsinki, Finland)
- Atte Hinkka  (University of Helsinki, Finland)
Thanks for your time!