VoG: Summarizing and Understanding Large Graphs

Danai Koutra
Jilles Vreeken
U Kang
Christos Faloutsos

HDMS 2014, Athens, Greece
[Originally presented at SDM’14, Philadelphia, PA]
Problem Definition

Given: a graph
Find: a succinct summary with possibly overlapping subgraphs
≈ important graph structures
Why graph summarization?

Visualization

Graph Understanding

Guiding attention
Application:
Wiki Controversial Article

I don’t see anything! 😞

Nodes: wiki editors
Edges: co-edited
Wiki Controversial Article

**Stars:**
- admins,
- bots,
- heavy users

**Bipartite cores:** edit wars
- Kiev vs. Kyiv
- vandals vs. admins

**Nodes:** wiki editors
**Edges:** co-edited
Main Ideas

Idea 1: Use a graph vocabulary:

Idea 2: Shortest lossless description

→ optimal compression (MDL)
Minimum Description Length

\[ \min L(M) + L(D|M) \]

\[ a_1 x + a_0 \]

\[ a_{10} x^{10} + a_9 x^9 + \ldots + a_0 \{ \} \]

simple & good explanations
Formally:
Minimum Graph Description

Given: - a graph $G$ with adjacency matrix $A$
       - vocabulary $\Omega$

Find: model $M$
      s.t. $\min L(G,M) = \min \{ L(M) + L(E) \}$
VoG: Overview

Step 1

Step 2

$\text{argmin}$

$\approx$?

$\approx$

Danai Koutra (CMU)
VoG: Overview

Step 3

Pick best (with some criterion)

Summary

Danai Koutra (CMU)
VOG is near-linear on # edges of the input graph.
VoG: Step-by-Step
We need candidate structures...

... How can we get them?
Step 1
Graph Decomposition

*Could use:*

ANY (overlapping) graph decomposition method

SlashBurn

(U METIS U spectral clustering U modularity...*)
SlashBurn-based Graph Decomposition

- **Slash** top-k hubs, **burn** edges
- Repeat on the remaining GCC

Notice that the structures *can overlap*!

[SlashBurn: U Kang and Christos Faloutsos. ICDM’11]
We got candidate structures.

Now, how can we ‘label’ them?
Step 2
Structure Labeling

\[ \approx ? \]

\[ \argmin \]

\[ \approx \]

Danai Koutra (CMU)
2.1 Structure Representation

hub?  

“best” node split?  

“best” node ordering?  

missing edges?
2.1 Structure Representation

**Hub**: top-deg node

**Spokes**: the rest

\[
\text{Hub: top-deg node}
\]

\[
\text{Spokes: the rest}
\]

\[
L(N|st|−1) + \log n + \log(\ ) + L(E+) + L(E-)
\]

\[
\text{Hub ID} n−1 | st |−1 \text{ spokes IDs extra missing Errors}
\]

**Star structure** + Errors

Danai Koutra (CMU)
2.1 Structure Representation

Max bipartite graph: **NP-hard**

**Heuristic**: Belief Propagation with heterophily for node classification (blue/red)

Bipartite graph structure + Errors

Danai Koutra (CMU)
Longest path: **NP-hard**

**Heuristic**: BFS + local search
2.2 Structure Labeling

\[ \text{argmin} \approx ? \]

MDL Cost
\[ L(m) + L(e) \]

Danai Koutra (CMU)
Step 3
Summary Assembly

Q: Which structures $S$ to pick?

A: Those that

$$\min_{S} L(G, M)$$
Summary Encoding Cost

\[ \min_{S} L(G,M) \]

\[ L(G, M) = 2^{|S|} \text{ combinations!} \]

for each structure
its encoding length
its type
its connectivity

# of structures
# of structures per type
its encoding length

\[ \text{DETAILS} \]

Danai Koutra (CMU)
Step 3: Summary Assembly

How to build the graph summary?
Step 3: Summary Assembly

Summary
Concepts

\[
\text{Savings} = \text{# bits as noise} - \text{# bits as structure}
\]
Step 3: Summary Assembly

(ii) Top-$k$

Summary

Danai Koutra (CMU)
Step 3: Summary Assembly

(iii) Greedy&Forget

Danai Koutra (CMU)
79,870 email accounts
288,364 emails

Danai Koutra (CMU)
Enron Summary

Top-3 Stars

klay@enron.com

jeff.skilling@enron.com

kenneth.lay@enron.com

---

Danai Koutra (CMU)
Enron Summary

Top Near Bipartite Core

Commenters CC’ed

“Affair”

organizers
participants
Ski excursion
VoG: summary

- Before:
  - Aggregate graph of non-overlapping supernodes with unknown properties
- Now:
  - Summary focuses only on important possibly-overlapping structures
  - With known graph-theoretic properties (beyond cliques)

Sense-making + visualization

VoG: Summarizing and Understanding Large Graphs.
Koutra, Kang, Vreeken, Faloutsos. SDM ’14.
www.cs.cmu.edu/~dkoutra/SRC/vog.tar
Thank you! Questions?

www.cs.cmu.edu/~dkoutra/pub.htm

danai@cs.cmu.edu