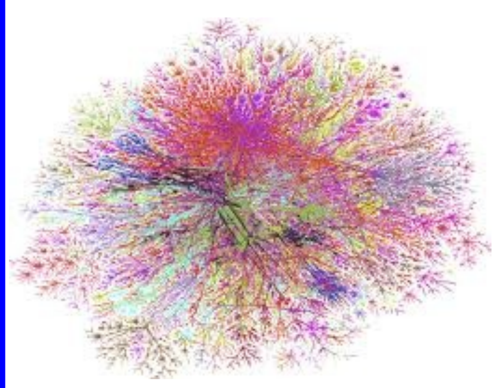
Perseus3: Visualizing and Interactively Mining Large-Scale Graphs

Di Jin[1], Ticha Sethapakdi[2], Danai Koutra[1], Christos Faloutsos[2]

[1] *University of Michigan, Ann Arbor* [2] *Carnegie Mellon University*

Our work:
Rich types of graph **summarization** and **interactive** subgraph visualization



Q1. How can we summarize large graphs of different **types** (unipartite or bipartite, directed or undirected)?

Q2. How to find specific **anomalous patterns** in large graphs effectively?

Q3. How do we achieve the above targets efficiently?

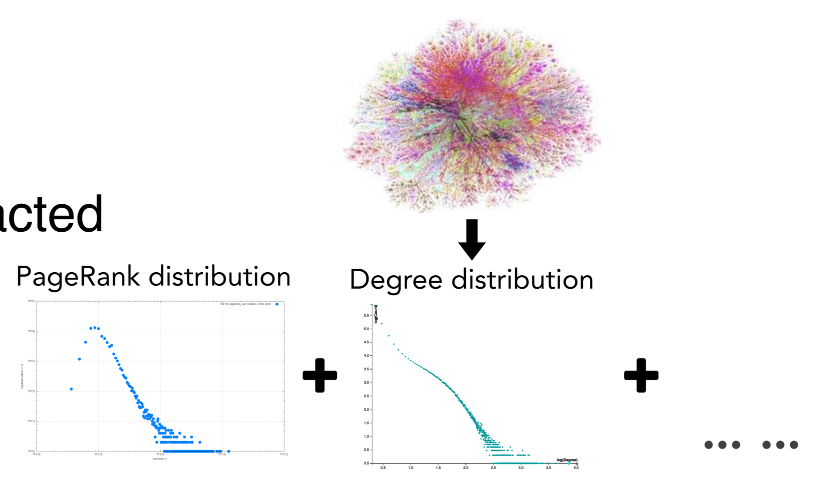
Solution 1: Graph Summarization

Problem Definition

Given a graph & type
Find comprehensive interactive and coupled graph statistics

Main Ideas

1. Multiple graph statistics are extracted and visualized.

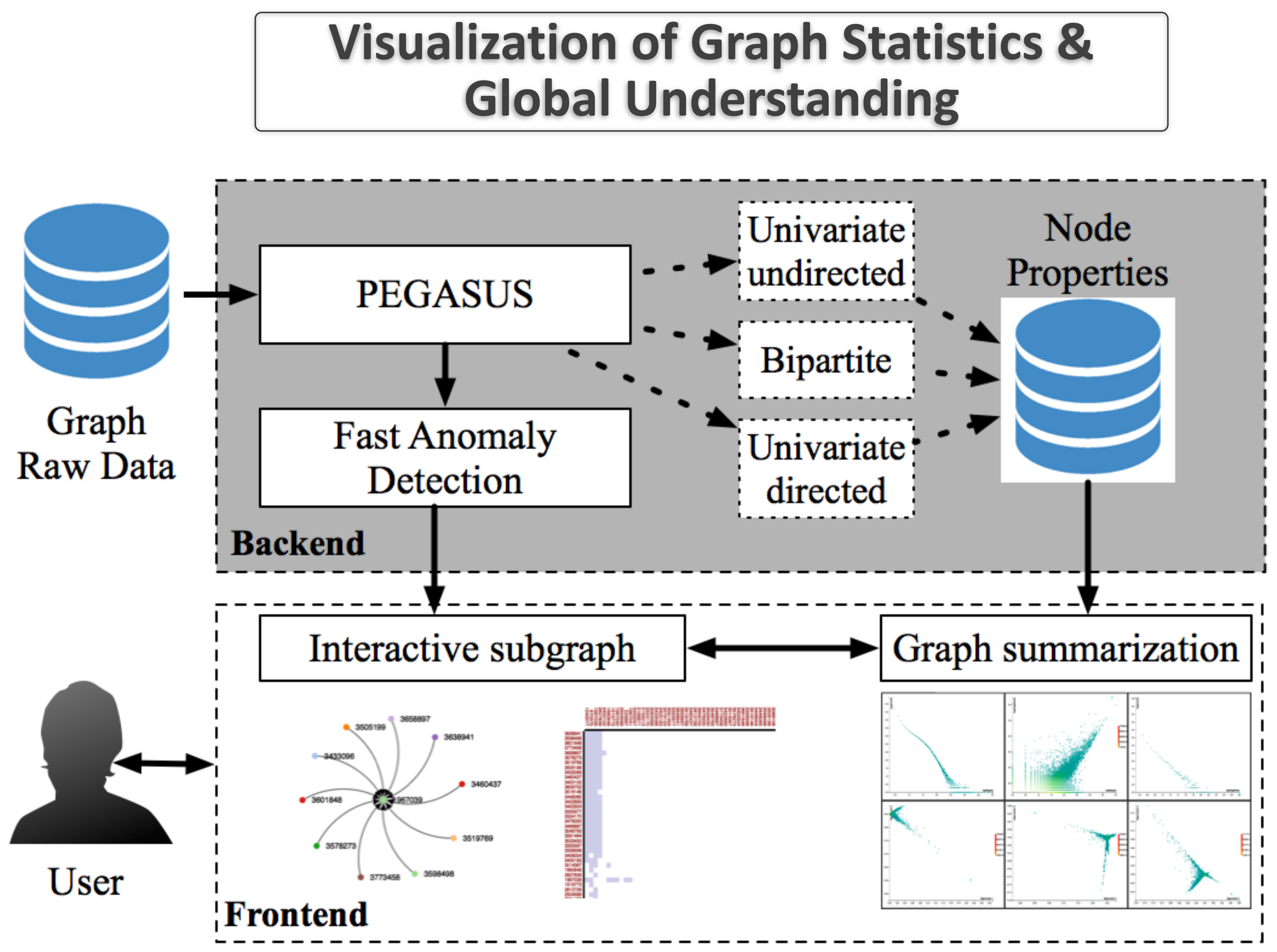


2. Various statistics are selected based on the type of the graph.

Graph type	Statistics
Unipartite + undirected	Total degree, PageRank, 1^{st}, 2^{nd}, 3^{rd} and 4^{th} eigenvector
Bipartite + directed	In degree, 1^{st}, 2^{nd} V vector (V1, V2), out degree, 1^{st}, 2^{nd} U vector (U1, U2)
Unipartite + directed	In degree, V1, V2 vector, out degree, U1, U2 vector

Table 1: Statistics visualized for each type of graph

Workflow

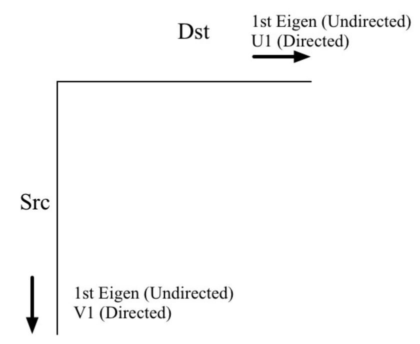


Solution 2: Interactive subgraph visualization

Problem Definition

Given a node in the graph
Find adjacency matrix for a specific node and 1-hop neighbors (aka. egonet subgraph)

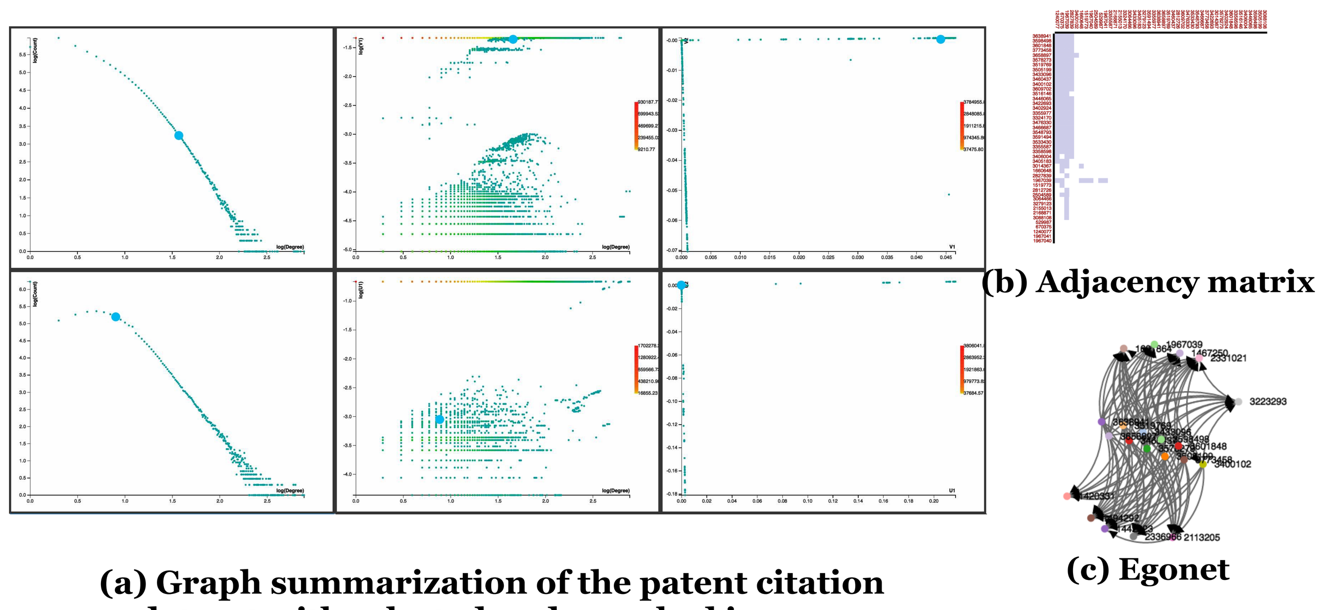
Main Ideas



1. Edges are stored in the database, corresponding nodes are ordered according to the graph type.
2. For bipartite graphs: For bipartite graphs, Local Sensitivity Hashing (LSH) is performed to find similar nodes based on common neighbors.
3. Nodes in the adjacency matrix are linked to dots in coupled distributions of graph statistics.

Comparison

Dataset: patent citation*, a directed graph containing 3,774,768 unique patents and 16,518,948 directed citations among them.



(a) Graph summarization of the patent citation dataset with selected node marked in cyan

(b) Adjacency matrix

(c) Egonet

* https://snap.stanford.edu/data/cit-Patents.html

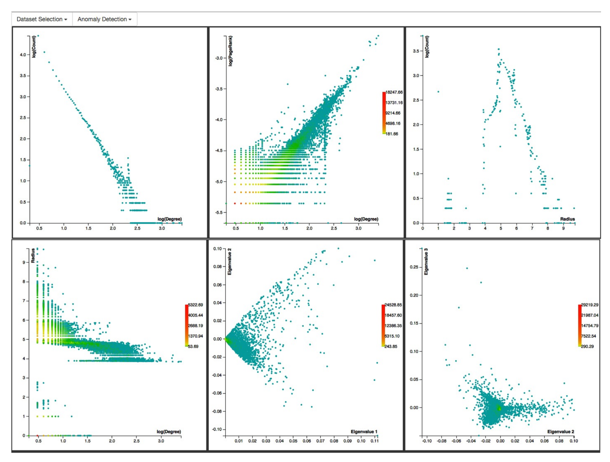
Solution 3: Heatmap representation

Problem Definition

Given the distribution of graph statistics
Find representation for better scalability

Main Idea

Points with identical graph statistics are aggregated in the distribution plots of graph properties and backend database to **a)** reduce storage, **b)** reduce burden to display and **c)** achieve 20x response time savings.



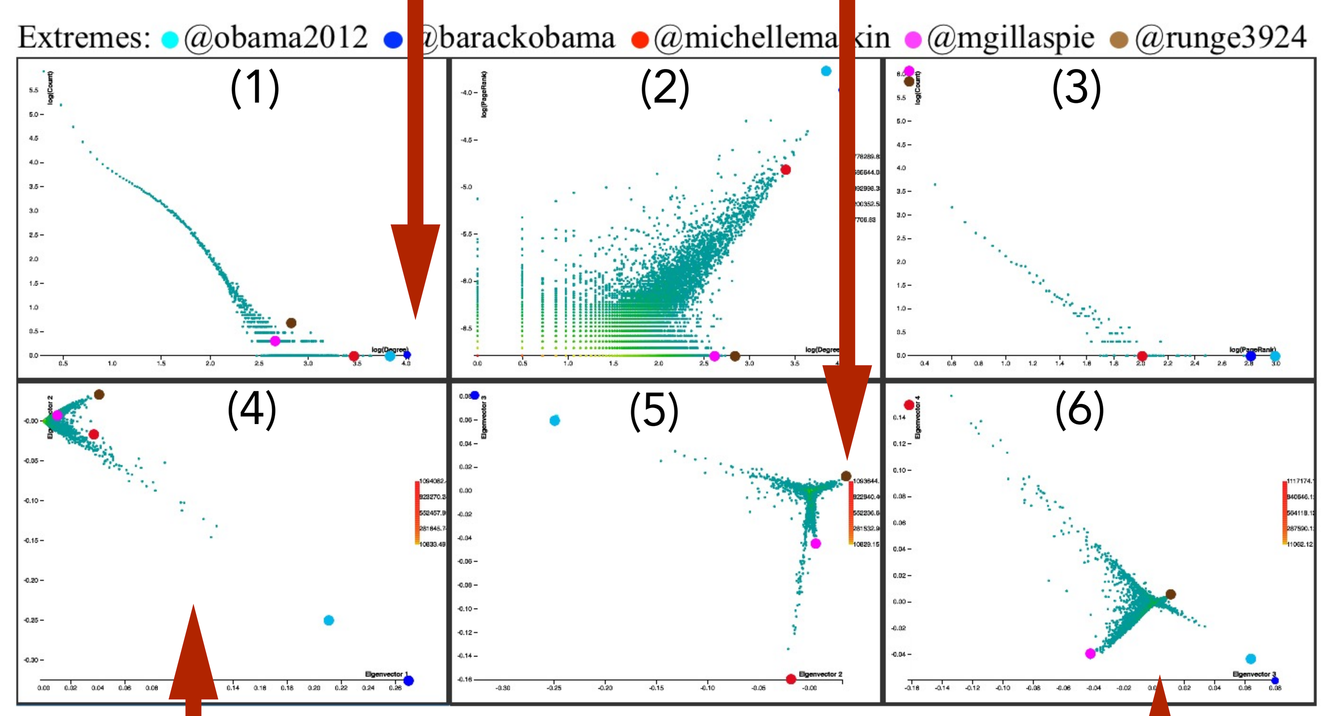
Applications

Undirected graph:

Dataset: One week in 2012 US presidential election period, containing 126,628 accounts and 4,191,918 tweets related to this topic.

Observation: Extremes in the eigenvector plots

1. The blue node is close to the cyan in every distribution. They turn out to be two accounts relevant to the same person.
2. The suspicious accounts such as @runge3924, form totally different communities from most users in the graph.



Graph summarization of the Twitter dataset with accounts of interest (marked in colors). PERSEUS3 helps spot at least 4 groups / spokes. Blue and cyan: President Obama (democrat); red: Michelle Malkin (conservative commentator); pink: mgillaspie (tea partier) and brown: runge3924 (suspicious account).

3. Retweeting behaviors can be revealed by eigenvector distributions. In plot (4), there are two spikes with extremes @runge3924 and @barackobama. @runge3924 has 1237 retweets but gets no retweeted while @barackobama retweets and gets retweeted the most.
4. Real users in Twitter tend to interact with people sharing the same interests thus forming communities with different topics. Since users of this dataset mainly focus on politics, different political communities are detected.

References

- PEGASUS: U. Kang, C. Tsourakakis, and C. Faloutsos. Pegasus: A peta-scale graph mining system - implementation and observations. ICDM, 2009.
- PERSEUS: Koutra, D., Jin, D., Ning, Y., & Faloutsos, C. (2015). Perseus: An Interactive Large-Scale Graph Mining and Visualization Tool. Proceedings of the VLDB Endowment, 8(12).