Graph Theory and Algorithms

Ph.D. Course – Marco Viviani

Graph Compression and Summarization
(April 29, 2021 / 15:00-17:00)

Graph Summarization

Intro and Taxonomy of Approaches

From: https://arxiv.org/abs/1612.04883

Definition and relationship with Compression

- **Graph summarization** methods leverage compression to find a smaller representation of the input graph, while discovering **structural patterns**.

- Graph summarization and compression are related.
  - In graph summarization, although compression is the means, finding the absolutely smallest representation of the graph is not the end goal.
    - The patterns that are being unearthed during the process may lead to suboptimal compression.
  - In graph compression, the goal is to compress the input graph as much as possible to minimize storage space, irrespective of patterns.

From: https://arxiv.org/abs/1612.04883

Benefits

- **Reduction**: volume and storage.
- **Speedup**: graph algorithms & queries.
- **Interactive analysis**
- **Noise elimination**

From: https://arxiv.org/abs/1612.04883

Applications

- **Clustering, Classification, Community detection**
- **Model order selection in matrix factorization**
- **Outlier detection, pattern set mining**
- **Finding sources of infection in large graphs**
- **Understanding selected nodes in graphs**

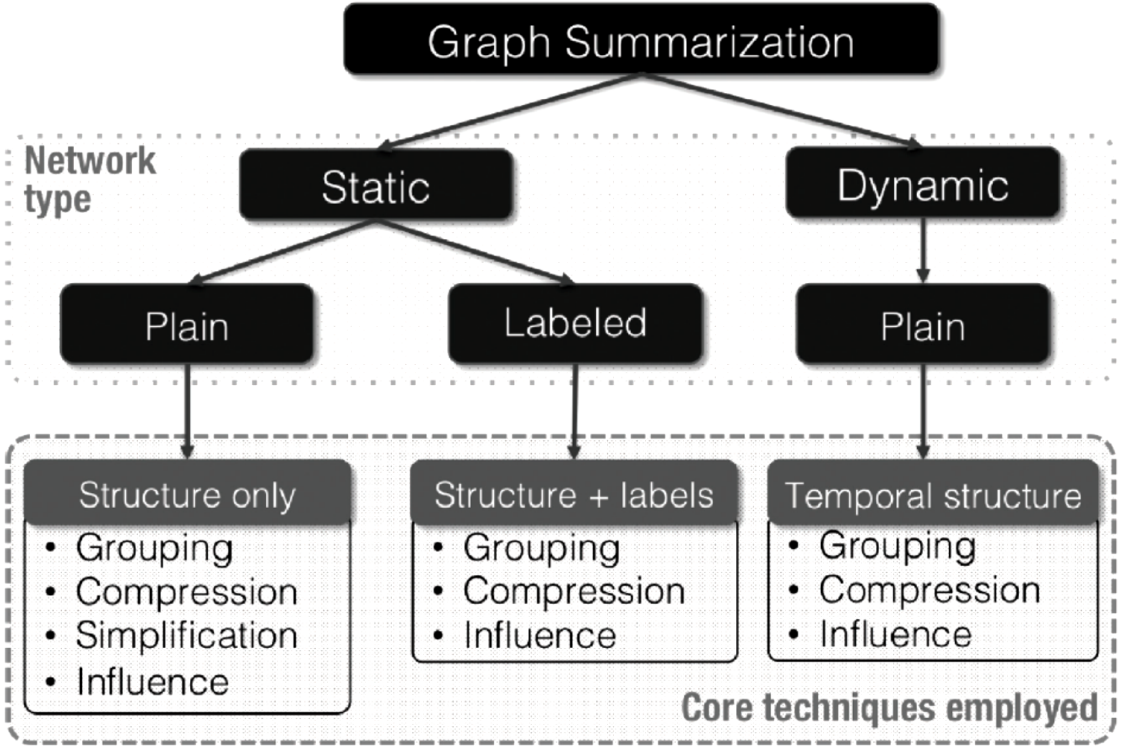
From: https://arxiv.org/abs/1612.04883

Challenges

- **Data volume**
- **Complexity of data**
- **Definition of «interestingness»**
- **Evaluation**
- **Change over time**

From: https://arxiv.org/abs/1612.04883

Graph Summarization Algorithms



From: https://arxiv.org/abs/1612.04883

Types of Graph Summaries

A **possible taxonomy** of graph summarization approaches can be divided into:

- Methods based on the **input data** handled.
- Methods based on the **core techniques** employed.

From: https://arxiv.org/abs/1612.04883

Input: Static or Dynamic

- Most summarization methods operate on **static networks**, leveraging graph structure (links), and, if available, the node/edge attributes.
- Despite the prevalence of large dynamic networks, only recent research efforts address their efficient summarization.
- In some cases, static methods are adapted to handle dynamic networks seen as series of static snapshots.

From: https://arxiv.org/abs/1612.04883

Input: Homogeneous or Heterogeneous

- The most well-studied instance in graph summarization, and graph mining more generally, is the **homogeneous graph** with one entity and one link type.

- However, some approaches apply to heterogeneous graphs by treating various types of nodes (e.g., students, instructors) and relations between them (e.g., teacher, friends, classmates) differently.

From: https://arxiv.org/abs/1612.04883

Core Technique: Grouping- or Aggregation-based

- This is the **most popular technique**.
- Some **node-grouping methods** recursively aggregate nodes into "supernodes" based on an application-dependent optimization function, which can be based on structure and/or attributes.
  - Others employ existing clustering techniques and map each densely connected cluster to a supernode.
- **Edge-grouping methods** aggregate edges into compressor or virtual nodes.

From: https://arxiv.org/abs/1612.04883

Core Technique: Bit-compression-based

- This approach, a **common technique in data summarization**, minimizes the number of bits needed to describe the input graph via its summary.
- Some methods are **lossless** and can perfectly reconstruct the original graph from the summary.
- Others are **lossy**, compromising recovery accuracy for space savings.

From: https://arxiv.org/abs/1612.04883

Core Technique: Simplification-based

- These methods streamline an input graph by **removing less "important" nodes or edges**, resulting in a simplified (or sparsified) graph.
- How to consider the concept of "impotance"?
  - Domain-/application-dependent.

From: https://arxiv.org/abs/1612.04883

Core Technique: Influence-based

- These approaches aim to discover a **high-level description of the influence propagation** in large-scale graphs.
- Techniques in this category formulate the summarization problem as an optimization process in which some quantity related to information influence is maintained.

From: https://arxiv.org/abs/1612.04883

What about the Output?

Output: Summary type

- A **supergraph**, which consists of supernodes or collections of original nodes, and superedges between them.
- A **simplified/sparsified graph**, which has fewer nodes and/or edges than the original network.
- A **list of (static or temporal) structures** or **influence propagations**, which are seen independently instead of in the form of a single summary graph.

From: https://arxiv.org/abs/1612.04883

What about the Output? … Cont'd

Output: Summary type

- **Flat**, with nodes simply grouped into supernodes.
- **Hierarchical**, with multiple levels of abstraction.

From: https://arxiv.org/abs/1612.04883

What about the Output? … Cont'd

Output: Non-overlapping or overlapping nodes

- In its simplest form, a summary is **non-overlapping**: each original node belongs only to one summary element (e.g., supernode, subgraph).
- Conversely, **overlapping** summaries, where a node may belong to multiple elements, can capture complex inherent data relationships but may also complicate interpretation and visualization.

4

Approaches

Graph Summarization

From: https://arxiv.org/abs/1612.04883

Grouping-Based Methods

- **Grouping-based (aggragation-based) methods** are among the most popular techniques for summarization.

- We distinguish grouping-based graph summarization methods into two main categories:
  - Node-grouping,
  - Edge-grouping.

From: https://arxiv.org/abs/1612.04883

Node-Grouping Methods

- Some approaches **recursively aggregate nodes** into supernodes, connected via superedges, based on an application-dependent optimization function.
- Others employ existing **clustering techniques** to find clusters that then map to supernodes.

From: https://arxiv.org/abs/1612.04883

Node Clustering-based Methods

- Although **the goal of clustering is not graph summarization**, the outputs of clustering algorithms can be easily converted to non-application-specific summaries.

- A small representation of the input graph can be obtained by
  - Mapping all the nodes that belong to the same cluster/community to a supernode.
  - Linking them with superedges with weight equal to the sum of the cross-cluster edges or else the sum of the weights of the original edges.

From: https://arxiv.org/abs/1612.04883

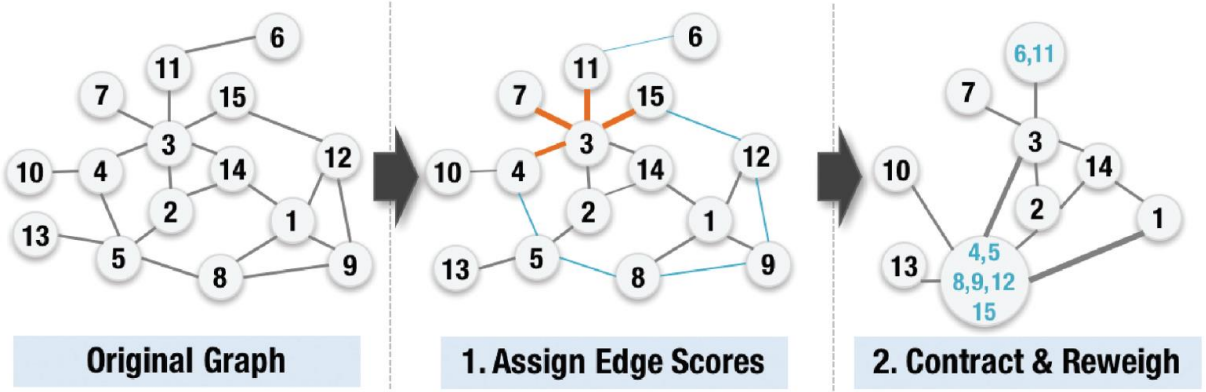
Node Clustering-based Methods … Cont'd

- Although the clustering output can be viewed as a summary graph, a **fundamental difference** from tailored summarization techniques is that:
  - Summarization groups nodes that are linked to the rest of the graph in a similar way.
  - Clustering methods simply group densely connected nodes.

From: https://arxiv.org/abs/1612.04883

Node Aggregation-based Methods

- **Aggregation-based methods** focus on the aggregation of the vertices of the graph. This aggregation is done using several strategies.



From: https://arxiv.org/abs/1612.04883

Node Aggregation-based Methods … Cont'd

- Some methods aims at **minimizing some version of the approximation or reconstruction error**.
  - GraSS (LeFevre and Terzi, 2010)
    - Normalized reconstructed error

$$E = \frac{1}{|V|^2} \sum_{i \in V} \sum_{j \in V} |\bar{A}(i,j) - A(i,j)|$$

  - Riondato et al. (2014)
    - l_p reconstruction error
    - The p-norm of

$$A - \bar{A}$$

Further info about norms

LeFevre, K., & Terzi, E. (2010, April). GraSS: Graph structure summarization. In Proceedings of the 2010 SIAM International Conference on Data Mining (pp. 454-465). Society for Industrial and Applied Mathematics.

Riondato, M., García-Soriano, D., & Bonchi, F. (2017). Graph summarization with quality guarantees. Data mining and knowledge discovery, 31(2), 314-349.

From: https://arxiv.org/abs/1612.04883

Toivonen et al. (2011)

- Focus on **compressing graphs with edge weights**, proposing to merge nodes with similar relationships to other entities (structurally equivalent nodes) such that approximation error is minimized, and compression is maximized.

- In merging nodes to obtain a compressed graph, the algorithm maintains either edge weights or strengths of connections of up to a certain number of hops.
  - In the simplest version of the solution, each superedge is assigned the mean weight of all edges it represents.
  - In the generalized version, the best path between any two nodes is "approximately equally good" in the compressed graph and original graphs, but the paths do not have to be the same.
    - The definition of path "goodness" is data and application dependent.

Toivonen, H., Zhou, F., Hartikainen, A., & Hinkka, A. (2011, August). Compression of weighted graphs. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 965-973).

From: https://arxiv.org/abs/1612.04883

Other Node-grouping Approaches

- Other node-grouping approaches seek **summaries that maintain specific properties of the original graph**.

- CoarseNet (Purohit et al., 2014)
  - The summarization problem is formulated as a minimization of the change in the first eigenvalue λ_1 between the adjacency matrices of the summary and the original graph.
- CoSum (Zhu et al. 2016)
  - It jointly condenses vertices into a supernode consisting of nodes of the same type with high similarity, and creates superedges that connect supernodes according to the original links between their constituent nodes.

Purohit, M., Prakash, B. A., Kang, C., Zhang, Y., & Subrahmanian, V. S. (2014, August). Fast influence-based coarsening for large networks. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 1296-1305).

Zhu, L., Ghasemi-Gol, M., Szekely, P., Galstyan, A., & Knoblock, C. A. (2016, October). Unsupervised entity resolution on multi-type graphs. In International semantic web conference (pp. 649-667). Springer, Cham.

From: https://arxiv.org/abs/1612.04883

Edge-Grouping Methods

- **Edge-grouping methods** aggregate edges into compressor or virtual nodes to reduce the number of edges in a graph in either a lossless or lossy way.

- Note that in this section, **“compression” does not refer to bit-level optimization**, as in the following section but rather to the process of replacing a set of edges with a node.

From: https://arxiv.org/abs/1612.04883

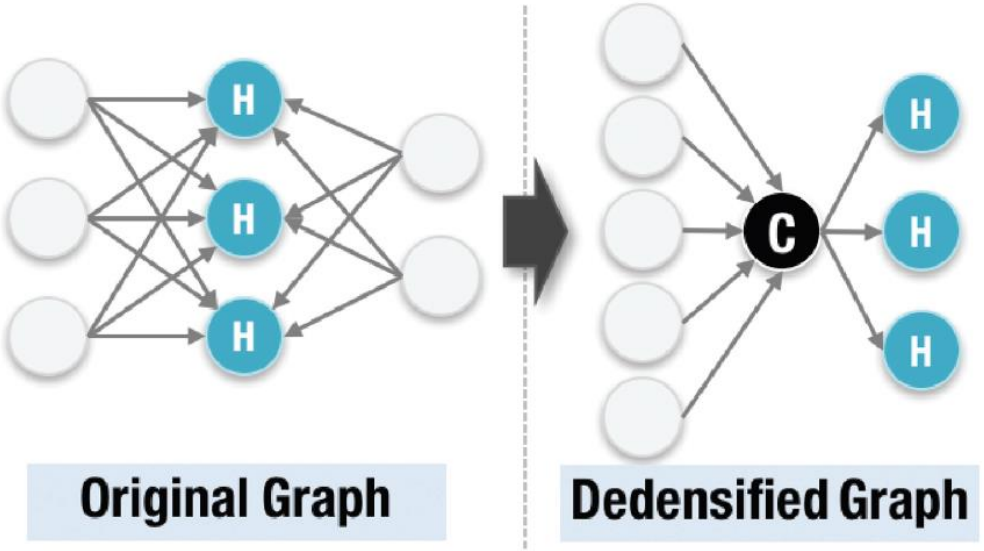
Graph Dedensification (Maccioni and Abadi 2016)

- An edge-grouping method that **compresses neighborhoods around high-degree nodes**.
- Following the **assumption that high-degree nodes are surrounded by redundant information** that can be synthesized and eliminated, the authors introduce “compressor nodes”, which represent common connections high-degree nodes.
- **Dedensification** only occurs when every node has at most one outgoing edge to a compressor node, and every high-degree node has incoming edges coming only from a compressor node.

Maccioni, A., & Abadi, D. J. (2016, August). Scalable pattern matching over compressed graphs via dedensification. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1755-1764).

From: https://arxiv.org/abs/1612.04883

Graph Dedensification (Maccioni and Abadi 2016) … Cont'd



From: https://arxiv.org/abs/1612.04883

Bit Compression-Based Methods

- **Bit compression** is a common technique in data mining.
- In graph summarization, the goal of these approaches is to **minimize the number of bits needed to describe the input graph**, where the summary consists of a model for the input graph and its unmodeled parts.
- The graph summary or model is significantly smaller than the original graph and often reveals various structural patterns, like bipartite subgraphs, that enhance understanding of the original graph structure.

From: https://arxiv.org/abs/1612.04883

Bit Compression-Based Methods … Cont'd

- As previously mentioned, **some of these approaches primarily use compression** and **secondary grouping techniques**.
- However, some **others aim solely to compress a given graph** without necessarily creating a graph summary or finding comprehensible graph structures → Graph compression!

From: https://arxiv.org/abs/1612.04883

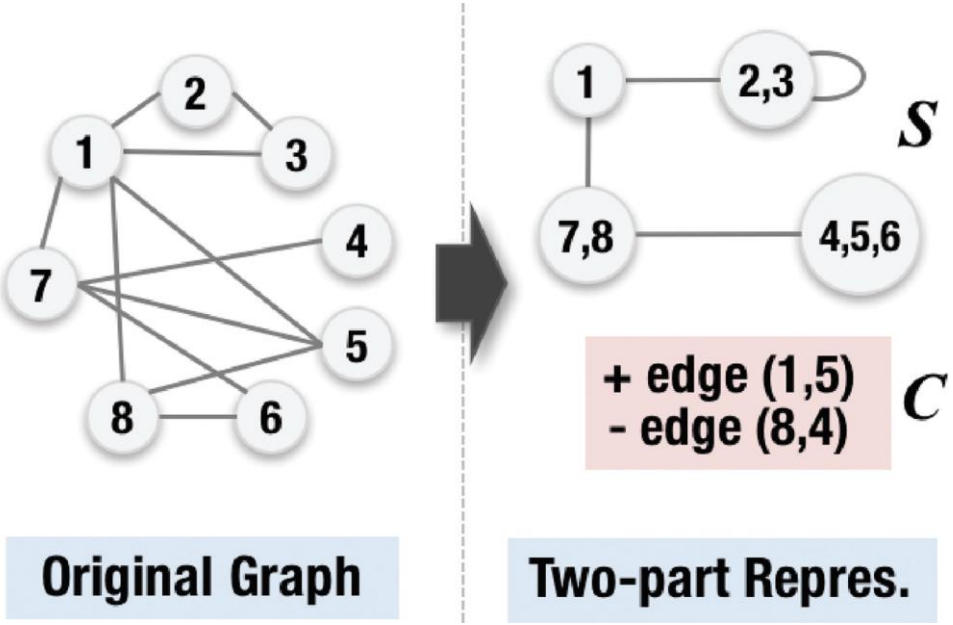
Two-part MDL (Navlakha et al., 2008)

- This representation, obtained by **aggregating nodes in the summary generation**, consists of a graph summary S and a set of corrections C.
- The summary is an aggregate graph in which each node corresponds to a set of nodes in G, and each edge represents the edges between all pairs of nodes in the two sets.
- The correction term specifies the list of edge-corrections that must be applied to the summary to exactly recreate G.

Navlakha, S., Rastogi, R., & Shrivastava, N. (2008, June). Graph summarization with bounded error. In Proceedings of the 2008 ACM SIGMOD international conference on Management of data (pp. 419-432).

From: https://arxiv.org/abs/1612.04883

Two-part MDL (Navlakha et al., 2008) … Cont'd

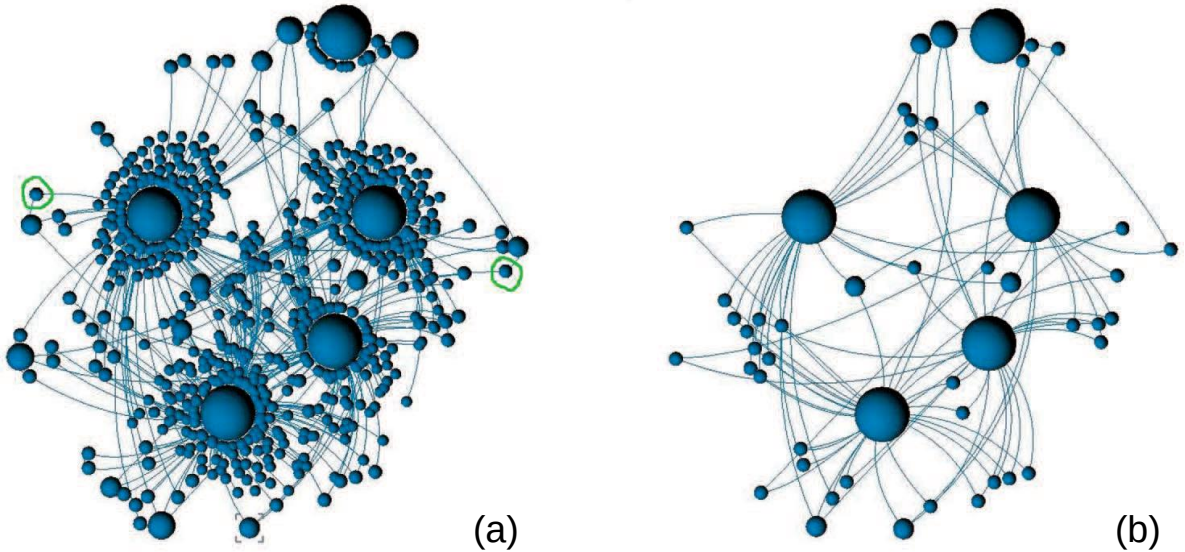


From: https://arxiv.org/abs/1612.04883

Simplification-Based Methods

- **Simplification-based summarization** methods streamline the original graph by **removing less "important" nodes or edges**, resulting in a simplified (or sparsified) graph.
- As opposed to supergraphs, here the summary graph consists of a **subset of the original nodes and/or edges**.
- In addition to simplification-based summarization methods, some existing graph algorithms have the potential for simplification-based summarization, such as:
  - Sparsification
  - Sampling
  - Sketching

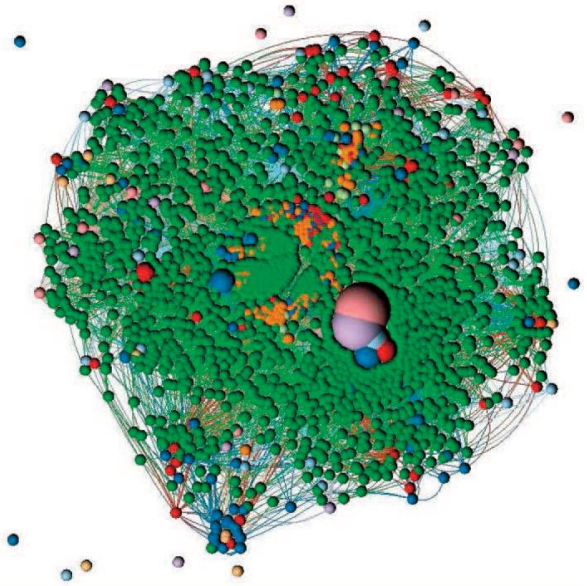
OntoVis (Shen et al., 2006)



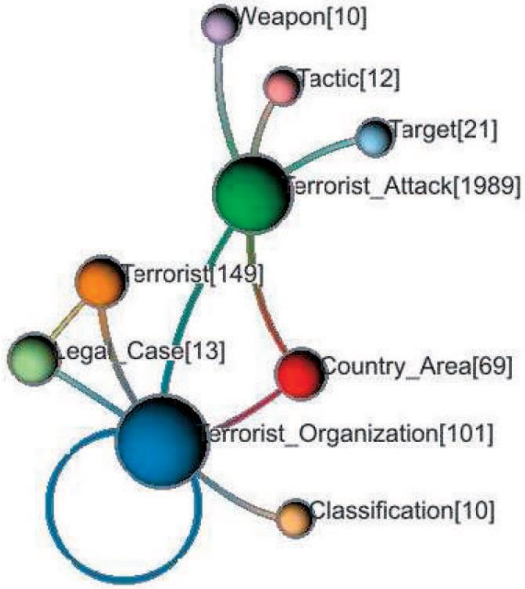
- (a) Original network has many one-degree nodes and duplicate paths
- (b) Structural abstraction after removing one-degree nodes and duplicate paths.

Shen, Z. et al. (2006). Visual analysis of large heterogeneous social networks by semantic and structural abstraction. IEEE transactions on visualization and computer graphics, 12(6), 1427-1439.

OntoVis (Shen et al., 2006) … Cont'd



Visualization of the Entire Terrorism Network. There are 2,374 nodes and 8,767 links.



Ontology Graph of the Terrorism Network. There are nine node types including terrorist organization, classification, terrorist, legal case, country/area, attack, attack target, weapon, and tactic.

From: https://arxiv.org/abs/1612.04883

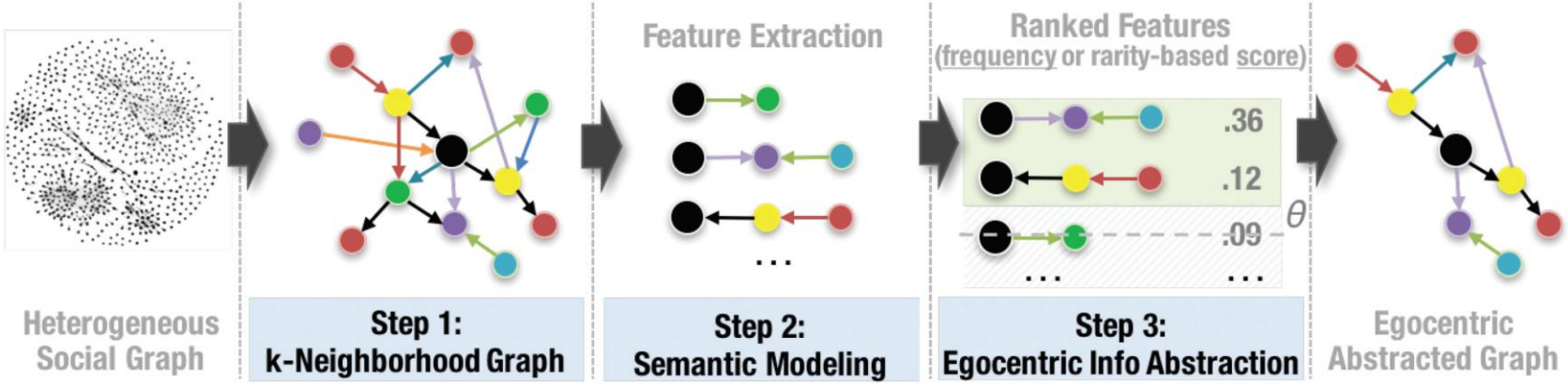
Egocentric Abstraction (Li and Lin, 2009)

- A **four-step unsupervised algorithm for egocentric information abstraction of heterogeneous social networks** using edge filtering (instead of node filtering)
  1. During the semantic modeling step, features (or else linear combinations of relations or path-based patterns) are automatically selected and extracted according to the surrounding network substructure (k-hop neighborhoods).
  2. The statistical dependency is measured between the features per ego node.
  3. During the egocentric information abstraction step, irrelevant information is removed by applying distilling criteria, such as keeping the most frequent or rare features.
  4. An egocentric abstracted graph is constructed incrementally on the remaining features, allowing the user to visualize the smaller resulting graph.

Li, C. T., & Lin, S. D. (2009, July). Egocentric information abstraction for heterogeneous social networks. In 2009 International Conference on Advances in Social Network Analysis and Mining (pp. 255-260). IEEE.

From: https://arxiv.org/abs/1612.04883

Egocentric Abstraction (Li and Lin, 2009) … Cont'd



From: https://arxiv.org/abs/1612.04883

Influence-Based Methods

- **Influence-based methods** seek to find a compact, high-level description of the influence dynamics in large-scale graphs to understand the patterns of influence propagation at a global level.
- Usually, such methods formulate graph summarization as an **optimization process** in which some quantity related to information influence is maintained.
- These summarization methods are scarce and have been mostly applied on social graphs, where important influence-related questions arise.

From: https://arxiv.org/abs/1612.04883

Community-level Social Influence (CSI) (Mehmood et al. 2013)

- **CSI focuses on summarizing social networks** via **information propagation** and **social influence analysis**.
- CSI **relies on existing clustering approaches**: it detects a set of communities using METIS (Karypis and Kumar 1999) and then finds their reciprocal influence by extending the Independent Cascade Model to communities instead of individual nodes.

From: https://arxiv.org/abs/1612.04883

Community-level Social Influence (CSI) … Cont'd

- Unlike other influence propagation approaches that find representative cascades for information diffusion, **CSI leads to a compact representation of the input network** where the nodes correspond to communities and the directed edges represent influence relationships.

- The **output of CSI** is different from grouping-based summarization techniques in which the superedges simply represent aggregate connections between the adjacent supernodes.

Possible Assignment

- Deepen a graph summarization approach that is based on the concept of maintaining the dynamics of influence within the graph.
- Critically discuss various models for evaluating the maintenance of influence within graphs.

Reference Literature

- Maneth, S., & Peternek, F. (2015). A survey on methods and systems for graph compression. arXiv preprint arXiv:1504.00616.
- Besta, M., & Hoefler, T. (2018). Survey and taxonomy of lossless graph compression and space-efficient graph representations. arXiv preprint arXiv:1806.01799.
- Liu, Y., Safavi, T., Dighe, A., & Koutra, D. (2018). Graph summarization methods and applications: A survey. ACM Computing Surveys (CSUR), 51(3), 1-34.
- Seo, H., Park, K., Han, Y., Kim, H., Umair, M., Khan, K. U., & Lee, Y. K. (2018). An effective graph summarization and compression technique for a large-scaled graph. The Journal of Supercomputing, 1-15.
- Kelly, T. (2020). Programming Workbench. Compressed Sparse Row Format for Representing Graphs. https://www.usenix.org/system/files/login/articles/login_winter20_16_kelly.pdf