Graph Mining Proposal

Yuwei Zhang
MCDS
CMU
yuweiz1@andrew.cmu.edu

Silun Wang
MCDS
CMU
silunw@andrew.cmu.edu

February 10, 2016

1 Survey

1.1 Papers read by Yuwei Zhang

Evaluating cooperation in communities with k-core structure [5]

- *Problem Definition*: The paper focuses on community detection and evalutation, which means dense connections among some of the nodes. The author make some novel changes to the k-core concepts including updating mtric for evaluating cohesiveness, assigning weights on the edges and other extended experimental evaluation.
- *Summary*:The original k-cores algorithm keeps deleting nodes whose degree are less than k and thus take the number of each set of vertices in the subgraph to do the evalution, which fails in the case where many co-authors have equal weight. This paper improves the method to define a co-authorship edge weight instead and recompute the evalution metrics with restrictions considered. In the expriment stage, testing on an unfiltered graph turns out to be extremely biased while on a filtered one(those co-author a lot) the results seem to be resonable. When weights graph method is applied, the extrem cases where k is too big are ignored and the algorithm gives better results.
- *Shortcomings*:There is no standard metrics to evaluate these algorithms/methods propsed in the paper. Also when we consider the graphs as social networks, where the relationship between two nodes are more than just co-author, for instance we have flollow, like, dislike, the weighted method should be furthur adjusted and it might be hard to derive the best weighting formula.

Vertexica: your relational friends for graph analytics [6]

- *Problem Definition*: To build a graph analysis tool, Vertexica, on top of a rdb that supports vertex-centric query interface. The system leverage the realational features and enable better graph analysis.

- *Summary*:Vertexica supports user-friendly and high-performance graph analysis by injecting data storage, query processing and query interfaces and supports various kinds of relational database. The system consists of four main components: physical storage to store data, coordinator as the center management driver, worker as the conntainer for the computation programs and vertex computation to process user queries. Vertexica also take several optimization techniques including: table union instead of table join, paralleling workers to work on multicores or multi machines, vertex batching to partition the table and create new tables other than update the orgin information to boost the performance. The paper also includes some use case demonstrations.
- *Shortcomings*:Hand-coded sql implementations give even better performance in the experiments of the paper. Is it possibel to furthur optimize the performance when using the user-friendly vertex-centric query interface?

Visual Exploaration of Collaboration Networks based on Graph Degeneracy [4]

- *Problem Definition*:To build a system that supports visual exploration of collaboration networks based on ranking of the nodes and filtering methods on the edges. It works on DBLP and is suitable for the large-scale networks.
- *Summary*:The idea of graph degeneracy is derived from the conceptsof k-cores, which is introduced in previous paper(the one just summarized). Basically in this system, it extracts the co-authorship graph using the algorithm described in the other paper using filtered weighted egde algorithm, and then partition the graph to f-cores based on the Trim process. Then comes the ranking, by repeatedly performing the Trim procedure to remove more vertexes and in the end stores in the relation databse for furthur query. The system can be useful to demonstrate bibliographic data.
- *Shortcomings*:For huge graphs, the k-core process may be extremly time-consuming. And it will take a long time for the system to reflect the updates in the graph.

1.2 Papers read by Silun Wang

The first paper was the Belief Propogation paper by Wolfgang Gatterbauer, Stephan Gunnemann, Danai Koutra, and Christos Faloutsos [3]

- *Problem Definition*: In big social networks, sometimes we need to infer the labels for particular nodes via transductive inference or semi-supervised learning. The classical belief propagation algorithm is widely used in such scenario, but it does not guarantee convergence in loopy graphs. In this paper, the authors propose Linearized Belief propagation and Single-pass belief propagation which are based on different restrictions and assumptions and much faster than BP.
- *Summary*: In a nutshell, LBP and SBP have the following advantages over BS:
  1. Have convergence guarantees
  2. Have closed-form solutions, thus reducing computational cost

3. Can be implemented on standardized SQL
4. SBP can be updated incrementally

LBP requires messages are normalized, thus the final belief matrix can be calculated via elegant matrix operations. SBP is based on the assumption that the impact of inference damps with length of paths. To obtain the final belief matrix, each node and each edge only need to be visited once.

- *Shortcomings*: The Daubechies wavelets require a wrap-around setting, which may lead to non-intuitive results.

The second paper was the k-core decomposition paper by Ignacio Alvarez-Hamelin, Luca DallAsta, Alain Barrat, and Alessandro [1]

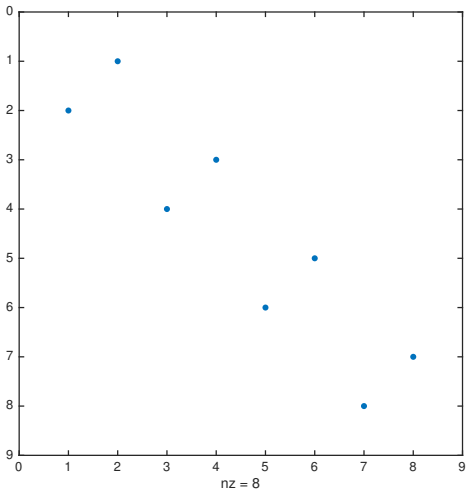
- *Problem Definition*: To visualize large complex networks is a big challenge, especially when you want clarity of graph and maintaining as much information as possible in the meantime. In this paper, the authors present an effective algorithm k-core decomposition to visualize large complex networks in 2D dimension.
- *Summary*: K-core decomposition introduces several terms: coreness, shell, cluster. It assigns each vertex a polar coordinate, thus visualizing a large complex network in 2D dimension while preserving relative hierarchical structures, connectivity and clustering properties, as well as interrelationship between hierarchies. Whats more, the overall time complexity is only linear as $O(n + e)$.
- *Shortcomings*: In order to obtain a readable layout, we need to tune several parameters. Can we learn these parameters automatically? Also, for huge networks, even a k-core decomposed graph seems to be nasty. Future work might need to combine nodes into a cluster and visualize a cluster via a simplified motif representation.

The third paper was the visualization paper by Cody Dunne and Ben Shneiderman [2]

- *Problem Definition*: Big data explosion results in huge and complex networks. In order to understand the relationship between entities and also individual attributes, traditional statistical charts are not applicable. Node-link diagrams are introduced and quickly excels among others. However, some node-link diagrams require relatively large screen space while containing little or repeated information, and optimization for the layout is NP hard. We need a more simplified visualization method which preserves important information.
- *Summary*: Authors of this paper on one hand defines three kinds of motifs: fan, connector and clique, on the other hand, presents an effective algorithm for motif detection with polynomial time complexity. After replacing the motifs with more representative glyphs, the graph requires much less screen space and layout effort. It helps us more easily understand the network and even discover some hidden relationships.
- *Shortcomings*: Users need to be trained for a short time to fully understand this new representation. It is ambiguous in choosing clique motifs because they often overlap with each other. Future work could present users with these overlaps and relative confidence on different partitions.

2 Unit Tests

2.1 Test Case I



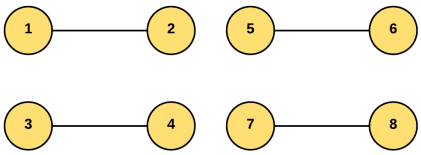
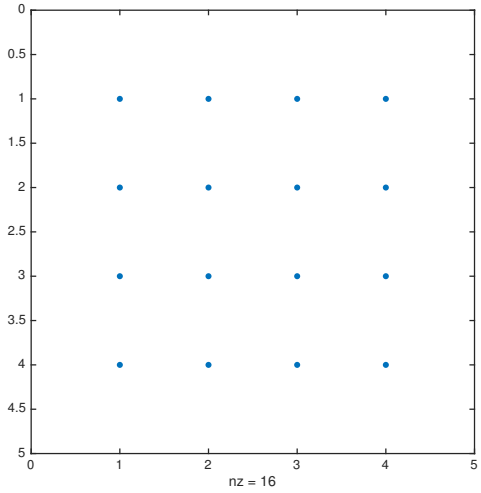


Figure 2: Graph

Figure 1: Adjacency Matrix

```
Output:
[(1, 1), (2, 1), (3, 1), (4, 1), (5, 1), (6, 1), (7, 1), (8, 1)]
```

2.2 Test Case II



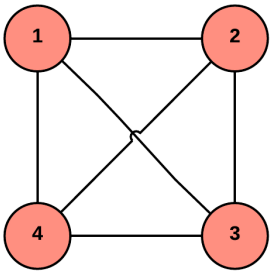


Figure 4: Graph

Figure 3: Adjacency Matrix

```
Output:
[(1, 3), (2, 3), (3, 3), (4, 3)]
```

2.3 Test Case III

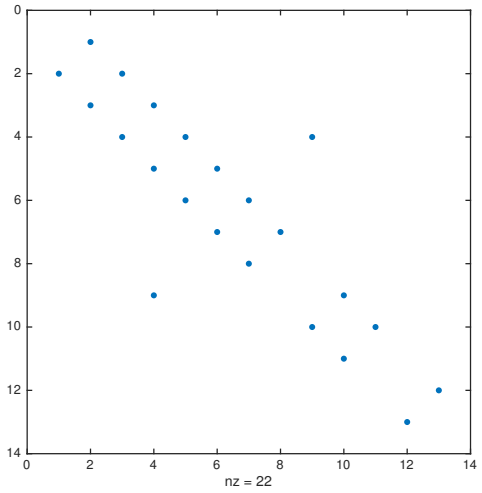


Figure 5: Adjacency Matrix

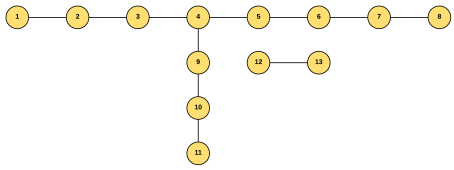


Figure 6: Graph

```
Output:
[(1, 1), (8, 1), (11, 1), (12, 1), (13, 1), (2, 1), (7, 1),
(10, 1), (3, 1), (6, 1), (9, 1), (5, 1), (4, 1)]
```

2.4 Test Case IV

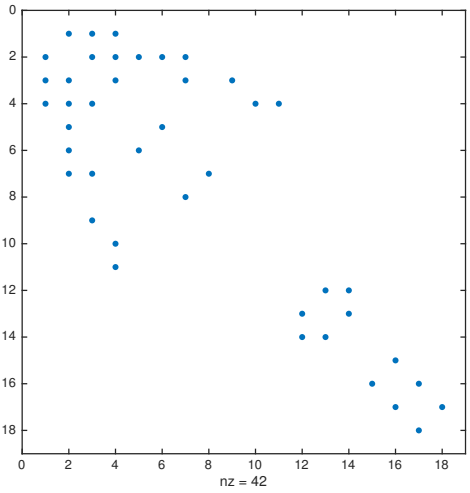


Figure 7: Adjacency Matrix

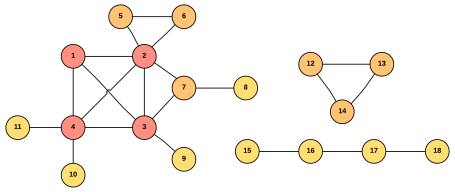


Figure 8: Graph

`Output:`
`[(8, 1), (9, 1), (10, 1), (11, 1), (15, 1), (16, 1), (17, 1), (5, 2), (6, 2), (12, 2), (13, 2), (14, 2), (7, 2), (4, 3), (1, 3), (2, 3), (3, 3)]`

2.5 Test Case V

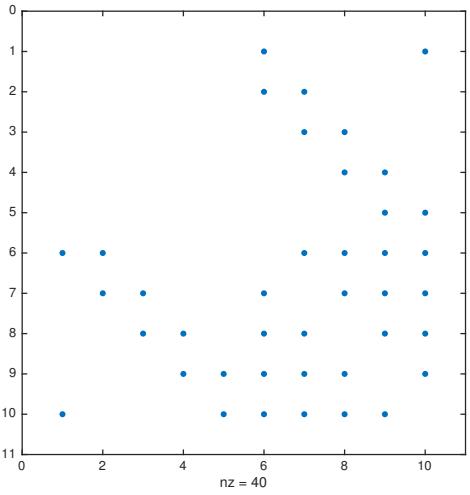


Figure 9: Adjacency Matrix

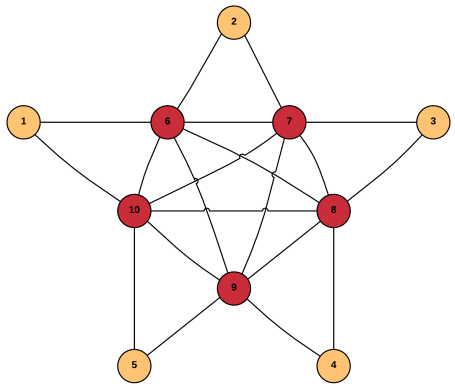


Figure 10: Graph

```
Output:
[(2, 2), (3, 2), (4, 2), (5, 2), (1, 2), (7, 4), (6, 4), (8, 4), (9, 4), (10, 4)]
```

3 Experiment Results

3.1 Core value

dataset	id = 0	id = 17	id = 9422	id = 18475	id = 27763
soc-Slashdot0811	43	43	43	43	15
soc-Epinions1	67	43	2	10	4

3.2 Degeneracy value

soc-Slashdot0811 : 55
soc-Epinions1 : 67

3.3 Core value distribution

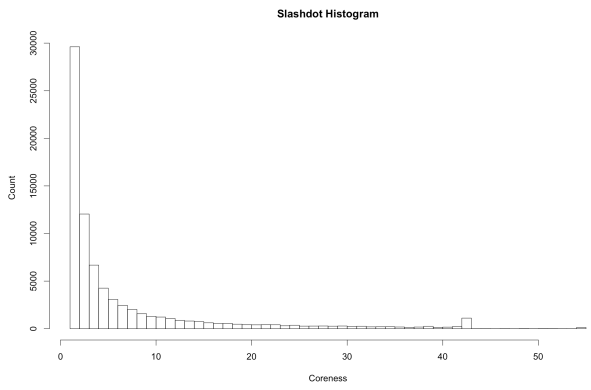


Figure 11: soc-Slashdot0811

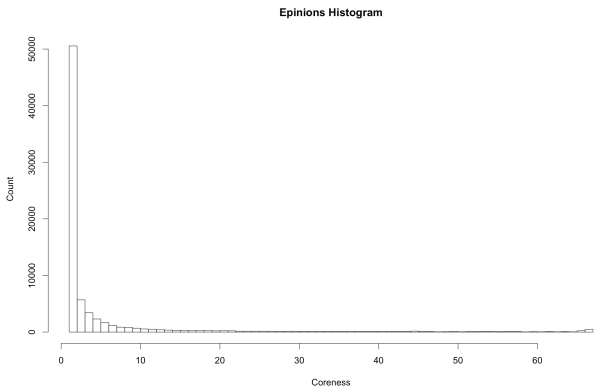


Figure 12: soc-Epinions1

References

[1] Ignacio Alvarez-Hamelin, Luca DallAsta, Alain Barrat, and Alessandro Vespignani. k-core decomposition: a tool for the visualization of large scale networks. In *arXiv.org*, 2005.

[2] Cody Dunne and Ben Shneiderman. Motif simplification: Improving network visualization readability with fan, connector, and clique glyphs. In *CHI*, 2013.

[3] Wolfgang Gatterbauer, Stephan Gunnemann, Danai Koutra, and Christos Faloutsos. Linearized and single-pass belief propagation. In *VLDB*, pages 581 – 592, 2015.

[4] Christos Giatsidis, Klaus Berberich, Dimitrios M Thilikos, and Michalis Vazirgiannis. Visual exploration of collaboration networks based on graph degeneracy. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1512–1515. ACM, 2012.

[5] Christos Giatsidis, Dimitrios M Thilikos, and Michalis Vazirgiannis. Evaluating cooperation in communities with the k-core structure. In *Advances in Social Networks Analysis and Mining (ASONAM), 2011 International Conference on*, pages 87–93. IEEE, 2011.

[6] Alekh Jindal, Praynaa Rawlani, Eugene Wu, Samuel Madden, Amol Deshpande, and Mike Stonebraker. Vertexica: your relational friend for graph analytics! *Proceedings of the VLDB Endowment*, 7(13):1669–1672, 2014.