Higher-Order HyperRank: PageRank Centrality by Hypergraph Motifs

Joshua Matt Turner
turnerjo@grinnell.edu
10/21/2020
CSC-395

1 Introduction

In many disciplines, it is useful to know which component of a complex system exerts the most influence on the operation and structure of that system. Thus, one of the key problems in network science is that of determining how important a node is based on network structure. This idea of node importance, and its computation, is called *centrality*. Since there are many ways that a node can be central, there are accordingly many centralities, many methods of assigning numbers to nodes such that more important nodes have higher numbers. Centralities have found applications in a wide variety of disciplines ranging from sociology to biology.

However, nearly all of these centralities are designed for the graph representation of networks, which, while undeniably useful, is nonetheless limited in its ability to express interactions between more than two entities. This limit presents significant problems for the expressiveness and accuracy of network modeling and associated analyses. To ameliorate this issue, network scientists have begun paying more attention to the *hypergraph* representation of complex networks, in which any number of entities can interact simultaneously. Despite their increasing popularity, hypergraph analyses, in particular hypergraph centralities, are lacking. Since hypergraphs are on the cutting edge of network science, it is of great importance to further develop tools for analyzing them.

This research introduces the Higher-Order HyperRank[1] centrality, a generalization of PageRank to hypergraphs that takes into account higher-order hypergraph structure. By considering not only the explicit relationships denoted by hyperedges, but also the implicit relationships suggested by interactions *among* hyperedges, this centrality provides a new perspective on node importance in hypergraphs.

Section 2 surveys past work in higher-order centralities. Section 3 lays the theoretical foundation for HR and sets up the brief experiments performed with the centrality. Section 4 discusses the results of those experiments and examines the shortcomings of HR. Section 5 provides directions for future research based on HR.

2 Previous Work

To understand the purpose of HR, it is necessary to situate it in the wider context of higher-order centralities. To say that a centrality is higher-order means that it takes larger structural features, such as network motifs, into account when ranking nodes.

An early yet powerful higher-order centrality is subgraph centrality, which ranks a node based on the number of closed walks that start and end in the node [5]. For example, the sequence $(v_1, v_2, v_4, v_5, v_1)$ is a closed walk on v_1. This centrality has a number of desirable features:

- it is easily calculated from graph spectra, since the number of closed k-walks on node i is $(\mathbf{A}^k)_{ii}$,
- it accounts for participation in network motifs, since shorter walks receive more weight, and real-world network motifs are small, and
- it is more discriminant between nodes than other centralities.

In a similar vein, [7] introduces a motif-based centrality, which, like the above, ranks nodes based on subgraph participation. However, this centrality only takes *specific* subgraphs – instances of a given network motif – into account when determining centrality. Thus, the more motif instances a node is part of, the more important it is. By paying attention to domain-specific "functional building blocks," this centrality outperforms many classical measures in identifying key proteins in PPI networks.

[1] abbrv. HR

[10] takes the idea of motif centrality one step further by integrating it with PageRank. This motif-based PageRank (MPR) is critical to the current project, as HR is the generalization of MPR to hypergraphs. The basic idea of MPR is to make nodes participating in instances a given motif more important. This choice is motivated by the fact that motifs are important structural features of networks; so, nodes that participate in them should be ranked as more important. The result is a centrality that outperforms standard PageRank on many networks. MPR will be covered in more detail in Section 3.

[9] provides another interesting higher-order centrality by ranking *groups of nodes*. The method presented defines the centrality of a subgraph as the proportion of network flows that the subgraph intercepts. Although the authors do not precisely clarify what this means, the centrality induces the eigenvector centrality over vertices, providing some intuition.

Higher-order centralities also include hypergraph centralities. [2] defines three eigenvector centralities on hypergraphs, and [6] uses hypergraph betweenness centrality on a gene network, finding that hypergraph betweenness outperforms graph betweenness.

3 Methods

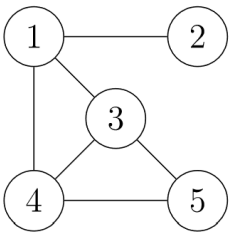
As mentioned above, this research attempts to generalize MPR to hypergraphs. So, before going any further, it is necessary to describe MPR in more detail, as well as introduce notions of hypergraph PageRank and hypergraph motifs.

3.1 Motif-based PageRank

Section 2 laid down the broad strokes of MPR as a modified PageRank that grants increased importance to motif-participating nodes. However, to motivate HR, we must look more closely at the underlying processes.

MPR consists of three steps: constructing a motif co-occurrence matrix, computing a weighted average of this matrix and the adjacency matrix $\mathbf{A}$, and performing PageRank on the result. The first step – constructing a motif co-occurrence matrix – consists of finding all instances of a given motif, and then creating a matrix $\mathbf{M}$ where $\mathbf{M}_{ij}$ is the number of motif instances in which nodes i and j co-occur.

The next step – computing a weighted average of the motif co-occurrence and adjacency matrices – involves choosing $\gamma \in [0, 1]$ and computing $\gamma\mathbf{A} + (1-\gamma)\mathbf{M}$. The below figure illustrates these two steps.



$$\begin{bmatrix} 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 2 & 1 \\ 1 & 0 & 2 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix} \qquad \begin{bmatrix} 0 & 0.6 & 1 & 1 & 0 \\ 0.6 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 1.4 & 1 \\ 1 & 0 & 1.4 & 0 & 1 \\ 0 & 0 & 1 & 1 & 0 \end{bmatrix}$$

Figure 1: A graph, its adjacency matrix, its triangle co-occurrence matrix, and a weighted average of the two ($\gamma = 0.6$). Notice that highly motif-participating edges have increased weight, while edges that do not participate in motifs have decreased weight.

The final step – PageRanking the result – is just as it sounds. Since edges that are part of more motif instances have more weight, their endvertices will have higher PageRank.[2]

The core concepts of the above centrality are motifs and PageRank. So, in order to generalize it to hypergraphs, we require suitable hypergraph analogues of the two. We first turn to hypergraph motifs.

3.2 Hypergraph Motifs

In contrast to graph motifs, which are subgraphs, hypergraph motifs (h-motifs) describe "the connectivity patterns of three connected hyperedges" [8]. In other words, an h-motif characterizes how three connected hyperedges

[2]They might not be strictly favored, but they will still have higher PageRank than would otherwise be given by the standard algorithm.

overlap (see Figure 2). This choice – to consider edge overlaps rather than subhypergraphs – is important, because it cuts away much of the variability intrinsic to hypergraphs, e.g. arbitrary edge and overlap sizes. By looking exclusively at three-way intersections, h-motifs allow for nuanced hypergraph analysis without having too much detail. Indeed, distinct domains exhibit different h-motifs, thus legitimizing the concept.

As made clear by the figure, an h-motif is defined by the emptiness of the 7 sections of its Venn diagram. While there are $2^7 = 128$ possibilities, 26 remain after accounting for symmetries and unrealizable motifs on distinct connected hyperedges.

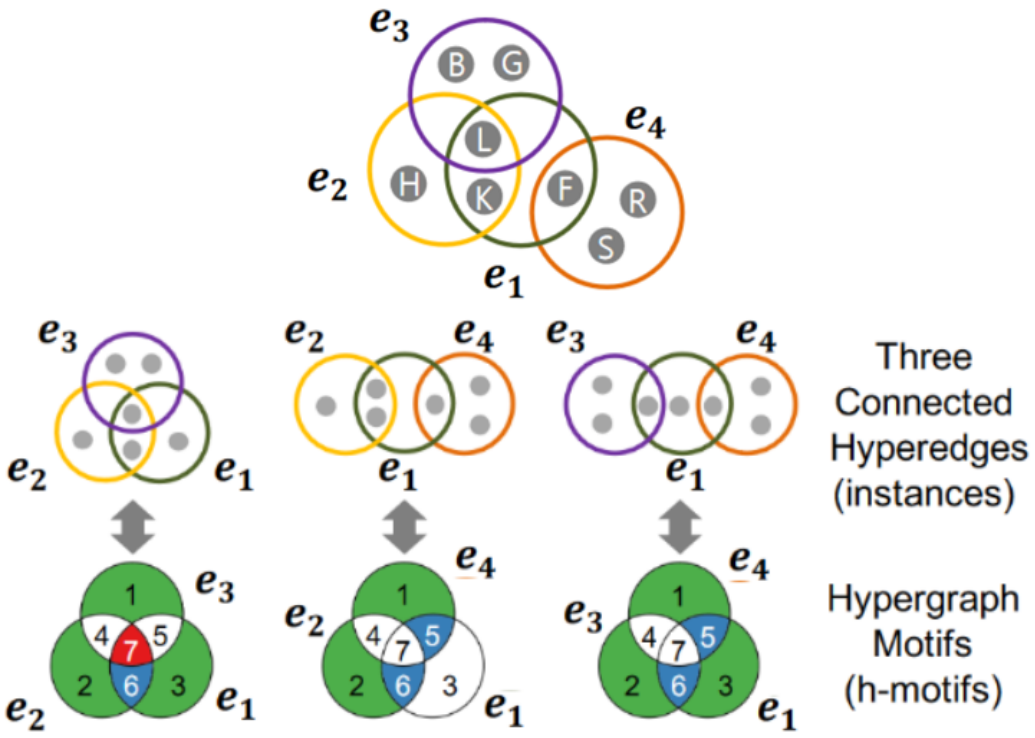


Figure 2: Instances of three h-motifs and their associated Venn diagrams. Base images obtained from [8].

Having discussed hypergraph motifs, we now turn to the other piece of the puzzle: hypergraph PageRank.

3.3 Hypergraph PageRank

The PageRank of a node is the probability that a random walk (which sometimes randomly teleports) is at that node at any particular time [4]. Thus, in order to extend PageRank to hypergraphs, all that is required is a notion of a hypergraph random walk. [1] provides one, via a transition matrix, as follows:[3]

$$P(u,v) = \begin{cases} \dfrac{1}{d(u)} \displaystyle\sum_{e \in \mathcal{E}(u) \cap \mathcal{E}(v)} \dfrac{1}{\delta(e) - 1} & u \neq v, \\ 0 & u = v \end{cases}$$

where $\mathcal{E}(u)$ is the set of hyperedges adjacent to vertex u, $d(u) = |\mathcal{E}(u)|$, and $\delta(e)$ is the size of hyperedge e. The intuition for the above formula is that the random walker first randomly selects an adjacent hyperedge and then randomly selects a node in that hyperedge.

Throughout the project, the running assumption was that a hypergraph random walk using this transition matrix would be equivalent to a graph random walk on the projected graph of the hypergraph. The projected graph $G = (V, E)$ of a hypergraph $H = (V', E')$ is the weighted graph where $V = V'$, $E = \{(u, v) : u, v \in e \text{ for some } e \in E'\}$, and $w(u, v) = |\mathcal{E}(u) \cap \mathcal{E}(v)|$. However, this assumption has since been discovered to be incorrect; rather, we should simply have $w(u, v) = P(u, v)$. PageRank on *that* weighted graph should be truly equivalent to hypergraph PageRank. We should expect random walks on the reweighted projected graph to be equivalent, since traveling between nodes is still pairwise. Although this error does not conceptually invalidate HR, the experimental results presented in Section 4 should not be taken as reflective of the process described in this section.

3.4 Higher-Order HyperRank

Finally, these two ideas come together to generalize MPR to hypergraphs. First comes the theory, then the implementation.

[3] A small modification has been made to prevent consecutive visits to the same node. A consequence is that 1-hyperedges are prohibited.

3.4.1 Theory

The goal of Higher-Order HyperRank is to give h-motif-participating nodes more PageRank than would be given by the standard algorithm. This stems from the idea that, since hypergraph motifs indicate important interactions in complex networks, nodes involved in those motifs must also be important.

In order to increase the ranking of h-motif-participating nodes, we leverage two observations:

1. Hypergraph PageRank is equivalent to PageRank on the reweighted projected graph with $w(u, v) = P(u, v)$.
2. Strengthening the weights in the projected graph between h-motif-participating nodes will, all else equal, increase their PageRank.

Thus, HR consists of the following steps, given a hypergraph $H = (V, E)$, a hypergraph motif m, $\gamma \in [0, 1]$, and $\alpha \in [0, 1)$:

1. Create the weighted adjacency matrix $\mathbf{A}$ such that $\mathbf{A}_{ij} = P(i, j)$
2. Create the motif co-occurrence matrix $\mathbf{M}$ such that
$$\mathbf{M}_{ij} = |\{e \in E : i, j \in e \text{ and } e \text{ appears in an instance of } m\}|$$
3. Normalize the matrices so that the largest entry in each matrix is 1.
4. Compute $\textsc{PageRank}(\gamma\mathbf{A} + (1 - \gamma)\mathbf{M},\ \alpha)$.

Normalizing the matrices ensures that one will not dominate the other due to a mere difference in scale. Also, note that step 4 is the MPR formula described in Section 3.1.

3.4.2 Implementation

The Julia implementation of Higher-Order HyperRank may be found at `https://github.com/joshua-matt/HO-HyperRank`. Although we will not discuss all of the code here, there are some features to note.

First, the hypergraph is represented as an incidence matrix $\mathbf{H}$, where $\mathbf{H}_{ij} = 1$ if vertex i is in edge j, and 0 otherwise. This representation, especially when using a sparse matrix, is much more efficient with space than the hyperedge list representation, and also allows for very easy calculation of the projected graph[4] ($\mathbf{H}\mathbf{H}'$) and the dual ($\mathbf{H}'$).

Second, the `MatrixNetworks` package is used for all graph-based computations, including PageRank.

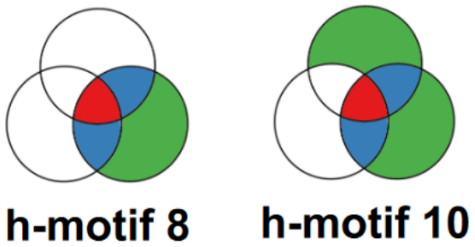
[4]That is, the projected graph where $w(u, v) = |\mathcal{E}(u) \cap \mathcal{E}(v)|$.

Third, and perhaps most important, is the current difficulty the program experiences for large datasets. This stems from, as one might expect, finding and classifying h-motif instances. The current implementation enumerates all unique triples of connected hyperedges and categorizes each as an instance of one of the 26 h-motifs.[5] Though significant improvements have been made compared to the beginning of the project, the speed still leaves much to be desired, especially since many hypergraph datasets are large. More will be said on this in Section 5.

3.5 Experiments

Despite computational limitations, the `Enron` dataset [3], at 10,885 hyperedges, is small enough for analysis. This dataset contains the correspondence patterns of 148 Enron employees, where each hyperedge contains the sender and recipients of a single email. In addition to its accommodating size, this dataset has also been widely analyzed, so there is a standard against which to compare experimental results.

Email datasets show a significant presence of h-motifs 8 and 10 [8], which are shown below.



The implication of h-motif 8 is that if two emails share members, then it is more likely that there is a third email which has all of them as members. This suggests a "clustering" of sorts, similar to triadic closure in graphs. Though this analysis does not concern itself with the chronology, it could either be that larger emails tend to "break off" into smaller emails containing subsets of the members, or that members of emails that share members tend to receive emails containing all members. H-motif 10 suggests a similar relationship

[5]Those including duplicate hyperedges, as well as any other invalid h-motifs, are also classified as such.

as h-motif 8, except that not all members of one of the smaller emails is included in the large email.

Given the significance of h-motif 8, it will be the motif used for HR on this dataset. It is worth noting that, before using HR, it is important to know which h-motifs are significant in the graph to be examined. Although h-motif 11 is a safe bet, since it is significantly present in all real hypergraphs studied in [8], it is nonetheless preferable to know the domain-specific motifs. An overview of domain-specific h-motifs may be found in Section 4.2 of [8].

In addition to performing HR on the dataset, we will also use other centralities for comparison. These are:

- The Clique, Z-, and H-eigenvector hypergraph centralities [2],
- Hypergraph PageRank, and
- MPR on the projected graph with a triangle motif.

As is typical, we use $\alpha = 0.85$ for the PageRank methods. We will also use $\gamma = 0.5$ for MPR and HR. Since these are the first experiments using HR, the best values for γ are unknown. In this ignorance, it seems safest to simply average the matrices.

4 Results

Before discussing experimental results, it is necessary to revisit the disclaimer of Section 3.3: due to the incorrect assumption that hypergraph PageRank is equivalent to PageRank on the projected graph, the results for hypergraph PageRank and HR **should not be taken seriously, since the results were obtained by executing a process different from what has been described here**. The results should be at least lightly correlated with those that would be obtained by the true process, since u and v sharing more hyperedges strictly increases $P(u, v)$; but, the degree of correlation is unknown. Thus, keeping their general invalidity in mind, we proceed to the results of the aforementioned experiment.

Looking at all centralities together (see Table 1), we see that both HPR (Hypergraph PageRank) and HR prioritize Jeff Dasovich, whereas the other centralities are unanimous in ranking Phillip Allen as the most important actor in the email network. Furthermore, Phillip Allen does not even appear in the top 10 for HPR or HR. It is here that we should remember the disclaimer. There is some degree of unanimity among all of them, as James

Email-Enron Rankings						
	CMEC	**ZEC**	**HEC**	**HPR**	**MPR**	**HR**
1	P. Allen	P. Allen	P. Allen	J. Dasovich	P.Allen	J. Dasovich
2	K. Presto	J. Steffes	K. Presto	R. Shapiro	M. Grigsby	R. Shapiro
3	J. Steffes	R. Sanders	J. Steffes	J. Steffes	S. Neal	J. Steffes
4	M. Grigsby	S. Kean	M. Swerzbin	T. Jones	J. Lavorato	T. Jones
5	S. Neal	R. Shapiro	M. Grigsby	M. Grigsby	S. Beck	S. Shackleton
6	M. Swerzbin	J. Dasovich	R. Badeer	J. Lavorato	B. Tycholiz	M. Grigsby
7	R. Badeer	M. Hain	M. Haedicke	M. Lenhard	M. Haedicke	J. Lavorato
8	M. Haedicke	R. Badeer	S. Neal	J. Arnold	J. Steffes	M. Heard
9	S. Kean	M. Swerzbin	S. Kean	R. Sanders	K. Holst	J. Arnold
10	F. Sturm	M. Haedicke	R. Shapiro	S. Shackleton	K. Presto	S. Bailey

Table 1: Employee rankings, where yellow cells indicate that the employee is not top 10 in any other centrality.

Steffes appears in the top 8 for all six centralities, and in the top 3 for all but MPR.

Now, comparing only HPR and HR, we first notice that the top four remain unchanged. Then, taking h-motif membership into account down-ranks most of the remaining: Grigsby, Lavorato, and Arnold all move down one spot, while Lenhard and Sanders leave the list entirely. However, Sara Shackleton flies up five spots, indicating her strong h-motif presence.

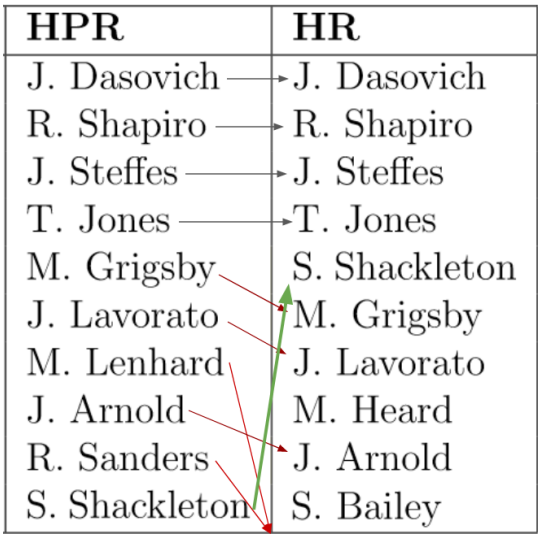


Figure 3: Comparison between HPR and HR

Analysis also yielded a couple interesting, unrelated facts: the set of 10 nodes that take part in the most triangles in the projected graph and the set of 10 nodes that take part in the most h-motifs are disjoint. This is very surprising, suggesting a potential error. Additionally, in the projected graph, only 500 out of the 20,000+ edges are between nodes that co-occur in an h-motif. This suggests a strong power law as to the number of h-motifs each nodes take part in.

5 Future Work

The exploration of HR provided in this project is limited, and there is much that could yet be done with this centrality.

First and foremost is an investigation into the significance of HR's rankings insofar as they differ from the standard hypergraph PageRank. Are these differences meaningful? Do these discrepancies, and does HR in general, capture important information about the network structure? Plainly, is HR actually useful?

Second, if the answer to the above is affirmative, is a firmer theoretical underpinning of HR. Are the intuitive foundations laid down in Section 3 rigorously supported? Is there a different characterization of hypergraph motifs that could yield meaningful results? What if we construct multiple motif co-occurrence matrices, each for different h-motifs, and weight those matrices differently in the weighted average, thus providing freedom to incorporate multiple motifs of varying importance? What else might be improved about this method? Furthermore, since PageRank is constant-sum – nodes gain importance at the expense of others – does a specific class of nodes tend to become less important as motif-participating nodes become more important? This last question is suitable for both MPR and HR.

Third is code optimization and correctness verification. The current runtime bottleneck is, as previously mentioned, the procedure that finds and classifies all hypergraph motif instances. The bottleneck of that procedure, in turn, is finding the intersections between all pairs of adjacent hyperedges. This inefficiency limits the size of hypergraph that can be comfortably analyzed to around 15,000 hyperedges. This is a problem, as many hypergraph datasets are considerably larger, with 100,000+ hyperedges. A modest speedup might be made by only considering one h-motif, and discarding any triples the moment one of its intersections does not match. However, if one

wishes to use HR with several different h-motifs, this would come at the cost of needing to search over the graph for each desired h-motif.

In addition, the motif-based PageRank method may not be correctly implemented, since very large changes occur as α departs from 1, with minimal change happening afterwards. On the other hand, HR displays slow, somewhat even change as α varies. Thus, additional experiments may be required in order to verify that the code performs the correct task.

Acknowledgement

The author would like to thank Nicole Eikmeier for her guidance in both the brainstorming and experimentation stages of the project.

References

[1] A. Bellaachia and M. Al-Dhelaan. "Random Walks in Hypergraph". en. In: (2013), p. 8.

[2] A. R. Benson. "Three hypergraph eigenvector centralities". In: *arXiv:1807.09644 [physics]* (Mar. 2019). arXiv: 1807.09644. URL: http://arxiv.org/abs/1807.09644.

[3] A. R. Benson et al. "Simplicial closure and higher-order link prediction". In: *Proceedings of the National Academy of Sciences* (2018). ISSN: 0027-8424. DOI: 10.1073/pnas.1800683115.

[4] S. Brin and L. Page. "The Anatomy of a Large-Scale Hypertextual Web Search Engine". In: *Seventh International World-Wide Web Conference (WWW 1998)*. 1998. URL: http://ilpubs.stanford.edu:8090/361/.

[5] E. Estrada and J. A. Rodríguez-Velázquez. "Subgraph centrality in complex networks". en. In: *Physical Review E* 71.5 (May 2005), p. 056103. ISSN: 1539-3755, 1550-2376. DOI: 10.1103/PhysRevE.71.056103. URL: https://link.aps.org/doi/10.1103/PhysRevE.71.056103.

[6] S. Feng et al. "Hypergraph Models of Biological Networks to Identify Genes Critical to Pathogenic Viral Response". In: *arXiv:2010.03068 [math, q-bio]* (Oct. 2020). arXiv: 2010.03068. URL: http://arxiv.org/abs/2010.03068.

[7] D. Koschützki, H. Schwöbbermeyer, and F. Schreiber. "Ranking of network elements based on functional substructures". en. In: *Journal of Theoretical Biology* 248.3 (Oct. 2007), pp. 471–479. ISSN: 00225193. DOI: 10.1016/j.jtbi.2007.05.038. URL: https://linkinghub.elsevier.com/retrieve/pii/S0022519307002780.

[8] G. Lee, J. Ko, and K. Shin. "Hypergraph Motifs: Concepts, Algorithms, and Discoveries". In: *Proceedings of the VLDB Endowment* 13.12 (Aug. 2020). arXiv: 2003.01853, pp. 2256–2269. ISSN: 2150-8097. DOI: 10.14778/3407790.3407823. URL: http://arxiv.org/abs/2003.01853.

[9] G. Pierre-Louis, t. l. w. o. i. a. n. w. Link to external site, and R. C. Wilson. "A centrality measure for cycles and subgraphs II". English. In: *Applied Network Science; Basel* 3.1 (Dec. 2018). Place: Basel, Netherlands, Basel Publisher: Springer Nature B.V. DOI: http://dx.doi.org.grinnell.idm.oclc.org/10.1007/s41109-018-0064-5. URL: http://search.proquest.com/docview/2427372109/abstract/5384BFF1998245AEPQ/2.

[10] H. Zhao et al. "Ranking Users in Social Networks With Higher-Order Structures". In: *AAAI*. 2018.