A Network-level View
of Author Influence

Henry Blanchette

May 1, 2019

1 Introduction

The data set I used for this study is the record of papers published in a selection of computer systems conferences (categorized by Frachtenburg) and a few control, non-computer systems conferences from 2017. The original format was the raw PDF's of all of these papers. During the summer of 2018, Frachtenburg and his research assistants extracted features from the PDF's including author names, paper titles, and bibliographies. The second layer of data embelleshment was matching authors with Google scholar accounts and personal survey results in order to attach author features such as email, institution, country, and gender. Finally, a third layer of data embelleshment was matching authors and papers to Semantic Scholar (*TODO:* citation) entries. Nearly all the papers and authors (*TODO:* percentage) from the original PDF's were found in Semantic Scholar. Importantly, these Semantic Scholar identities significantly contributed to author disambiguation and retrieved the papers from previous years that were cited by the 2017 papers.

The stated purpose for my work on and analysis of this data was to look for patterns in the authors' and papers' individual-level features (from the embelleshed data) and network-level features. I analyzed several different networks, focussing on the collaboration and citation connections between authors, papers, and conferences. I found that the author collaboration network was particularly interesting because it displayed some significant, meaninful and separate correlations between individual-level features and network-level features. These correlations were novel in the context of some more tradition *author reputation metrics*, which suggest that they measure an aspect of influence that is not well measured by the reputation metrics.

2 Background

The study of social networks in general has detailed how networks can give rise to novel, emergent and important features. These sorts of features include: diameter, connected components, bridges, distance, and centrality. I mainly focus on centrality, but the others (especially connected components

and bridges) will also play well into the later analyses as complements in explaining centrality distributions.

A node's centrality in a network can have several different flavors as will be described, but abstract, centrality is a how *influential* a node is on the rest of the network (where "influential" is in terms of whatever the edges in the network represent). There are a few specific measures of node centrality that have proven most useful for many network contexts. They are *degree*, *eigenvector*, *betweenness*, and *closeness* centrality.

Degree centrality is the simplest. The *degree* of a node is the number of edges that include the node (note that there can also concepts of in-degree and out-degree that only count edges of a certain orientation with respect to the given node). Then the degree centrality of a node is degined just to be the node's degree. Often, since centralities are considered at a network-level context, degree centrality is measured differently from degree only in being normalized over the degrees of the nodes of the entire network. Degree centrality is a measure of how prolifically connected a node is to the network, as just a raw count of how many connections it has. This measure is the most common used in network analysis, but it is also the most local of the centrality measures because it does not consider the arrangement of the network outside of each nodes' direct neighbors.

Eigenvector centrality is more complicated. The concept is that it weights the centrality of each node by the centralities of its neighbors, after starting from some base ranking of nodes by degree. In this way, it is closely related to PageRank (*TODO:* citation). Formally, the eigenvector centrality of the nodes in a network is given by the vector x that satisfies the equation

$$Ax = \lambda x$$

where A is the adjacency matrix of the network and λ is an eigenvector with non-negative entries (*TODO:* citation). This requirement yields that there is a unique solution up to scaling. Therefor, eigenvector centrality is only meaningfully considered when normalized over the eigenvector centralities of all the nodes in the network. Eigenvector centrality is a network-level-focussed measure because it weights heavily not just the direct neighbors of a node but also the neighbors of its neighbors and so on with diminishing effect.

Betweenness centrality measures how efficiently a node connects the network. A given node a's betweenness centrality is defined to be, over all pairs of nodes c, d in the a's connected component, the fraction of shortest paths between c, d that include a. Finally, the fraction is weighted by the size in nodes of a's connected component. Betweenness centrality is strongly connected to the concepts of connected components and bridges in a network. A connected component is a subset of the nodes of a network that have paths between any pair of them, but no paths to nodes outside the component. Bridges are nodes that connect two would-be (non-trivial) connected components. These bridges often have high betweenness because any path, and thus including the shortest ones, between nodes on opposite sides of the bridge must go through it.

Closeness centrality measures how close to the "center of mass" of the network a node is. Closeness of a node is defined to be the connected-component-weighted reciprocal of its *distance*, where the distance of a node is the mean length of the shortest paths between it and each other node in its connected component. In a geographical way, many networks arrange themselves with a sort of planar flatness, so closeness centrality can be seen as how close a node literally is to the geographical center of the network. Closeness is related to how involved a node is with the rest of its connected component as opposed to being involved with only a specific fringe (even if it is avidly involved which would lead to a high degree centrality).

All of these metrics are useful in the context of measuring "influence," but, of course, influence can be specified more clearly in specific situations. For example, scientific collaboration analysts of computer science fields have used degree centrality alng with correlations of node traits to explain why authors collaborate and how likely authors' with certain traits (e.g. gender, institution type) are to collaborate in the future [Ghiasi G.]. Along with degree, connections between nodes can be classified by other individual- and network-level features of the connected nodes [Ghiasi G].

More generally, the collaboration networks of scientific researhers have suggested that some specific features are important for how likely authors are to collaborate. These features include: co-authorship distance in the author collaboration network, geographical closeness, cultural closeness, and likeness of research institution type (e.g. university or indistry) [Knoke D, Yang S]. Some interesting results pointed to the conclusion that, perhaps, inter-disciplinary collaboration may be comparably important to intra-disciplinary collaboration [Knoke, Yang S].

3   Hypotheses

My first inquiry was into the topical relationships of the conferences of the papers into the data. Which conferences were most closely related to "computer systems" research, and how did the other distribute? The data also contained papers from conferences that were identified as outside of computer systems in order to but the topical categorization into scale. My hypothesis was that the typical intuitions about which computer systems conferences were most computer-systems-y would be right. Of all people the scientists should know what kind of work they are doing, right?

My second inquiry was into the significance of typical reputation measures for authors, such as hindex and i10index, at the network level of analysis. More specifically, I predicted that reputation would correlate with the network-level sense of centrality that would be captured by some of the centrality metrics detailed in section 2. The relationship between reputation and network-level node features should reveal something in regards to what the reputation metrics are and aren't good measures of.

4   Conference Citation Network

In the conference citation network, each node represents a 2017 conference and each directed edge represents a citation by a paper in the source node's conference of a paper in the target node's conference. Multiple with the same source and target combine into a single edge by summing weights (where weight is the number of citations that an edge represents). These citations can go back in time however, as the conferences included in the data repeat regularly over the last couple decades. Many citations were of conferences that were not present in the 2017 data; these were excluded from analysis.

TODO: conference citation network figure with labels, colors, sizes.

TODO: look back at this figure and make some judgements.

TODO: relate to first research question of how system-y certain conferences are.

5   Paper Collaboration Network

In the paper collaboration network, each node represents a paper from the 2017 data and each undirected edge represents an author that is shared by each of the connected node's papers. Multiple edges between the same two nodes combine into a single edge by summing weights (where weight is the number of shared authors that an edge represents).

As 1 illustrates, this network is very clean visually. It also gives a sense for the shape of the network. There is a distinct, largest connected component that contains the vast majority of the highly degree-centric nodes. *TODO:* number of connected components and distribution of sizes Outside of this there are only a few other notably-large components, and only one of which has notably-high degree-centric nodes (located in the bottom-right of figure 1). *TODO:* name some papers in this group

I found, however, that for the purpose of analyzing the features of authors, this network was sub-optimal. Lumping together authors by paper obscure what are likely important author attribuets that are heterogeneous among the co-authors of a particular paper. Taking these considerations into account, I went on to create analyze the network in section 6.

6   Author Collaboration Network

In the author collaboration network, each node represents an author and each undirected edge represents a paper from the 2017 data that the connected authors collaborated on. Multiple edges between the same nodes are combined into a single edge by summing weights (where the weight of an edge is the number of co-authored papers between the connected nodes' authors).

As is immediately obvious from figure 2, there is one particular group of nodes that is skewing the degree centrality of the whole network. This fully-connected clique of authors are those that authored the *TODO:* which paper is this? paper.

TODO: figures for all the other centraltiy metrics

TODO: distributions of centralities

TODO: authors features pairs matrix

7   Conclusions

7.1   Conferences' Thematic Relations to Computer Systems

TODO: reflect on observations in the conference section

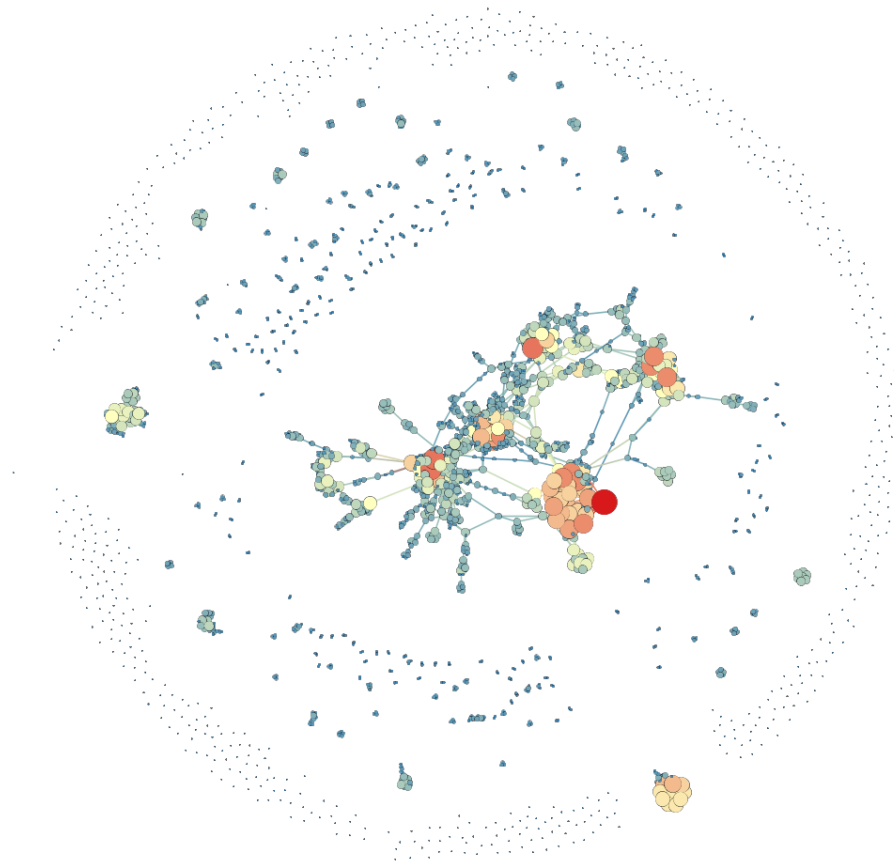## 7.2   Reputation and Centrality

References

Figure 1: The paper collaboration network where the size and color of nodes correspond to their degree centrality. The color scale is from blue (low centrality) to white (middling centrality) to red (high centrality).
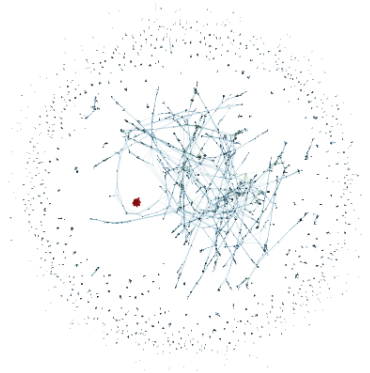
Figure 2: The author collaboration network where the size and color of nodes correspond to their degree centrality. The color scale is from blue (low centrality) to white (middling centrality) to red (high centrality).