Realistic Social Networks

To illustrate that different methods of modeling a social network affect results, I plan to use an approximation of a real-world social network. Because very large datasets of scientific metadata are now available though the Microsoft Academic Graph (MAG), Web of Science and others, we can now get pretty good estimations of who is listening to whom in science. So, to illustrate the above point that graph choice matters, I plan to create a more realistic graph from the MAG that can stand in for Zollman's idealized cycle and complete graphs. The goal here is not to find a perfect approximation, merely one that is more empirically motivated than Zollman's and results in radically different simulation results. If the conclusions of the simulation cannot be drawn from this graph structure, then the idealized graph structures can be seen as leading to conclusions that are unlikely to generalize to real, complex social structures.

Chosing Realistic Subgraphs

First, I chose to use the MAG as a starting point for creating empirically-motivated graph structures. The MAG includes detailed publication metadata, including titles, citations, authors, and fields of study for a dizzying number of academic works. Because there are so many types of data, picking what data type in the MAG becomes a node and which relations become edges is a critical part of creating a convincing approximation of a real social network. Furthermore, the overall dataset size is measured in hundreds of gigabytes and millions of works, meaning finding ways to limit the size of a network are critical because running a simulation on all or a large portion of the dataset is computationally intractable.

Before making decisions selecting node types, edge types and a way of selecting a subgraph, it is critical to define which aspect of social interaction I aim to capture in the first place. Because Zollman's model models scientists "communication of results", I intend my subgraph to approximate authors reading each others work. When an academic reads another academics work, information, positive or negative, about the publishing author's work and approach is transmitted to the reader. Thus, I wish to approximate the "readership" network of who is reading who because that network seems to most closely resemble the interaction of communicating experimental results that Zollman's model is intended to model.

Now, the MAG does not track readership because it is impossible to know for sure what authors have read what papers without surveiling and publishing the personal data of scientists, so we need some sort of data as a stand-in. For this purpose, I chose citation of a paper as a proxy for reading a paper. This assumes that the citing author has read and been influenced in some way by the cited paper. This isn't always true, in the case of a fluff citation added without any knowledge of the cited paper, I'd argue its a pretty good approximation as it is hard to find and cite a paper without gleaning any information from it. It is important to note that this does not assume cited papers are read or understood,

just that they've influenced the citing author in some way and thus constitute some sort of communication of information.

Now, that I've supported the idea of using citation to approximate transmission of results, let me more rigorously define how I create subgraphs. Because in Zollman makes authors nodes in a graph, the nodes in our subgraph are authors. Now, I draw an edge between two authors A and B if and only if A wrote a paper that cited a paper by B. This definition should capture pathways by which information in science is transmitted. Note that this definition results in a *directed* graph because A can write a paper citing B without B returning the favor. We could make an undirected graph from this same relation by assuming that when an author is cited, that auther might be aware or become aware of the citing author.

While this definition works pretty well, I needed a practical way of selecting which authors should be modeled from the vast array of authors in the MAG. To do this, I decided to choose a subfield, and connect only authors that have published in that field. Furthermore, because the MAG includes many types of publications from journal articles to books, I limited my search to journal articles that referenced at least one other paper and were cited at least once. In an ideal word with reliable data and fast simulations, I wouldn't limit this search, but this felt like a good way to make simulations tractable while ensuring all article entries I included were not junk or incomplete.

To select this subgraph, I loaded the MAG into the neo4j graph database and queried it to create subgraphs using the query defined below. Once the query finishes, I convered the results to a networkx graph in Python, which can save the graph in the GraphML format for later use in a simulation. Note: because the cited relation is directed as I mentioned above, the graph we save is directed.

```
def get_cited_relations(fieldOfStudyName):
    CITED_QUERY = """
        MATCH (p1:Paper)-[:IN_FIELD]->(parent:FieldsOfStudy{normalizedName: $fosName})
        WHERE p1.citationCount > 0 AND p1.referenceCount > 0 AND p1.docType = "Journal" WITH
        MATCH (p1:Paper)-[r:REFERENCES]->(p2:Paper)
        WHERE p2.citationCount > 0 AND p2.referenceCount > 0 AND p2.docType = "Journal" WITH
        MATCH (p2:Paper)-[:AUTHORED_BY]->(a2:Author)<-[:AUTHORED_BY]-(:Paper)-[:IN_FIELD]->(
        WITH p1, a2
        MATCH (p1:Paper)-[:AUTHORED_BY]->(a1:Author)
        RETURN DISTINCT a1, a2
        LIMIT 50000
    """
    with driver.session() as session:
        return session.run(CITED_QUERY, fosName=fieldOfStudyName).data()

def authors_to_graph(authorPairs):
    edge_list = [(p['a1']["normalizedName"], p['a2']['normalizedName']) for p in authorPairs
    G = nx.DiGraph()
```

```
    G.add_edges_from(edge_list)
    return nx.convert_node_labels_to_integers(G)
```

NOTE: I plan to choose more graphs, but right now querying the MAG for subgraphs can take quite a while and many fields I've run into have turned out to be quite big.

I picked the field of social epistemology (MAG normalized name "network epistemology") itself as it felt fitting to approximate and model the field itself. It is also a nice size with 799 authors with a paper published according to the criteria I defined above, meaning it was tractable both for neo4j to construct the graph and to run as a simulation. I queried and calculated the following stats:

```
rels = get_cited_relations("social epistemology")
g = authors_to_graph(rels)
print("Number nodes: {}".format(len(g.nodes())))
print("Number edges: {}".format(len(g.edges())))
print("Number strongly connected components: {}".format(nx.number_strongly_connected_compone
print("Number weakly connected components: {}".format(nx.number_weakly_connected_components(
nx.write_graphml(g, "social_epistemology.graphml")
```

```
Number nodes: 799
Number edges: 1999
Number strongly connected components: 723
Number weakly connected components: 78
```

Note that having so many strongly connected components and relatively few weakly connected components indicates that this citation relation is rarely reciprocated. When author A cites author B, author B rarely cites back or cites another author who cited A.