Empirically-Motivated Model Mechanism: A Case Study

Given the mechanistic interpretation I've given of Zollman's model in *The Epistemic Benefit of Transient Diversity*[1] and my account of how mechanistic models can be evaluated based on how well the model's mechanism represents that of the target's mechanism, I wish to demonstrate how viewing models in this way leads to more robust evaluation of a model. Because making the entire model more empirically-motivated would be a difficult undertaking, especially given the model includes models of poorly-understood phenomena like belief representations in people, I choose to focus on the model mechanism part Zollman focuses on: the network structure of scientists. This experimental section then has two distinct parts: developing a realistic and defensible model of network structure following from empirical data and verifying that new model by attempting to replicate Zollman's modeling results on these larger, more complex graphs. My primary goal with this experimental work is less to evaluate Zollman's model, but rather to present empirically-defined mechanism as another way to evaluate mechanistic models in terms of empirical distance, as discussed in previous sections. Viewed in opposition to Rosenstock et al.'s critique of Zollman's model on the grounds that the results don't hold outside a specific parameter range, my experimental work seeks to determine how the model performs on parameter values determined by empirical data, rather than mathematical idealizations. In short, if the model mechanism looks a lot like the target mechanism, then how does Zollman's model perform?

Part 1: Empirically-Motivated Social Network Models

Graph Terminology

First, I will define some background terminology that is critical to understanding both Zollman's models and my extensions of them.

A graph G, mathematically, is a tuple $G = (N, E)$ where N is a set of *nodes* and E is a set of tuples of vertices where $(n_i, n_j) \in E$ represents an *edge* or connection between node $n_i \in N$ and node $n_j \in N$. Nodes are often also referred to as vertices, however I adopt Zollman's use of "node" for clarity here. This structure of nodes and relations between nodes shows up in many fields such as model theory in modal logics, biological modeling of complex systems and social network modeling. Graphs are very adept at capturing relational structure in the real world, which brings up an important distinction between *graph* and *network*. When I say graph, I refer to the mathematical structure given above. A graph has no interpretation, it is merely some structure made of nodes and edges. A *network*, on the other hand, refers to a relational structure in the real world which can be modeled as a graph. For example, if we want to model the

[1] Zollman, "The Epistemic Benefit of Transient Diversity."

internet as a graph (simplistically), nodes might be computers and edges might be network links between them.

Graphs may be *directed* or *undirected*. In a directed graph, an edge represents a one-way connection where the edge (n_i, n_j) denotes a connection from n_i to n_j but not n_j to n_i. In an undirected graph, all edges are bidirectional such that (n_i, n_j) means both n_i connects to n_j and n_j connects to n_i. In practice, an undirected graph can be realized in a directed graph so long as the edge $(n_i, n_j) \in E$ implies the presence of the edge $(n_j, n_i) \in E$.

Because graph structure can be so rich, there are far more ways to quantify structure than can be detailed and deployed in a single paper. Instead, I will introduce a few fundamental properties that are essential to understanding Zollman's model here. A connected graph is one in which there exists a path connecting any two nodes in the graph, where path is defined as a sequence of nodes $n_1, \ldots, n_l$ such that for each n_i in the path, there exists an edge $(n_i, n_{i+1}) \in E$. In an undirected graph, the order of an edge is ignored. For directed graphs, there are two senses of connectivity *weak* and *strong*. Weak connections treat the directed graph as undirected (ignoring direction of edges) whereas strong connections do not.

To illustrate these definitions, consider the following directed graph $(N, E) = (\{1, 2, 3\}, \{(1, 2), (1, 3)\})$. The graph is weakly connected (and thus connected if undirected) because the node 1 is connected to node 2 via $(1, 2)$ and to node 3 via $(1, 3)$. Node 2 is connected to node 1 via $(1, 2)$ and to node 3 via $(1, 2)$ and $(1, 3)$. Node 3, similarly to node 2, is connected to node 1 via $(1, 3)$ and to node 3 via $(1, 3)$ and $(1, 2)$. Note that nodes 2 and 3 do no connect to any other node if edges are instead directed. Thus, this graph is weakly connected, but not strongly connected.

Finally, this definition of connectivity leads to the idea of *connected components* when combined with the the idea of a *subgraph*. A *subgraph* $G' = (N', E')$ of a graph $G = (N, E)$ is formed by a subset of nodes $N' \subseteq N$ and and a subset of edges $E' \subseteq E$ such that E' contains only edges where both endpoints are in N'. Now, a connected component is a subgraph of G such that G' is connected. We can further call a strongly connected G' a strongly connected component, and likewise for weakly connected components.

Finally, there are several types of ideal graphs that are often discussed and that Zollman uses as test cases in his paper. First, a *complete* graph is one in which every node is connected to every other node in a graph. Second, a *cycle* is a strongly connected graph in which every node is connected to exactly two neighbors such that all nodes are connected in one large ring. Finally a *wheel* is a strongly connected graph identical to a cycle, with the addition of a single central node which is connected to every other node.

Zollman's Communication Networks

While I will not rehash my framing of Zollman's model mechanistically in this section, I will spend a bit of time focusing on the technicalities of how the network structure part of the model operates. I'll begin with a discussion of the intended target of Zollman's model to clarify what a charitable interpretation of his work might be, then a discussion how this intention is represented in the actual model machinery.

Zollman discusses social networks, where he defined individuals as nodes and edges to be the "the communication of results from one to the other."[2] Furthermore, he posits that this relationship is symmetric, so one if individual A can view individual B's results, B can also view A's results. This definition can mean quite a few things in practice and will prove too broad to apply directly to measurable real-world behaviors. To demonstrate this, consider the following cases where results are communicated:

1. A scientist reads a published work, finds the results interesting and eventually cites that work in her own work building off of or criticizing the published work.
2. A scientist reads a published journal article, finds the results uninteresting and does not ever cite the work.
3. A scientist reads a news article about an a research work, is influenced by the high-level ideas, but never reads underlying academic work and thus does not cite it.
4. A scientist emails a friend in another lab for advice about starting a project and the friend reports the their results in the project area didn't look promising. Nothing is published, but info about results is transmitted.
5. A prominent scientist writes a blog or tweet reacting to a paper, influencing people's opinions of that paper without any published work to document it.
6. A scientist runs into another researcher at a conference and the two informally share ideas.

Beyond these, there are a multitude of other ways by which results may be communicated within a scientific community with varying degrees of impact and evidence associated with the transfer. Many of the more informal methods of communication would be very hard to measure or quantify on a large scale. Informal conversations and emails are rightly private and not available for public analysis and social media is an emerging form of communication for scientists which is not yet well-understood,[3] meaning these forms of communication of results are difficult to measure and quantify.

Citations, on the other hand, formally recognize the specific results of prior work which influenced the researcher. The APA's influential publication manual formalizes this influence-based understanding of citation, recommending that

[2] P. 25 Zollman.

[3] Collins, Shiffman, and Rock, "How Are Scientists Using Social Media in the Workplace?"

researchers "Cite the work of those individuals whose ideas, theories, or research have directly influenced your work."[4] These citations can ofter take on a more strategic rhetorical purpose, explicitly building on a paradigm of work started by another or criticizing that paradigm. This argumentative conception of citation's role in scientific community is characterized in greater depth by Bruno Latour in *Science in Action* as a part of his Actor-Network theory,[5] however, for our purposes here, we need not wade into the subtleties here because no matter how a citation is deployed in this sense, it trivially can be said to have influenced the author. Even if the author is just cursorily familiar with the work and is citing to criticize the approach, they display some awareness of the results found. Citations don't always live up to this intent, as authors may cite simply to help get by publication referees and in some cases journals even force authors to cite certain articles, forming "coercive citations."[6] Despite these degenerate cases, I argue it is fairly safe to assume that a majority of authors who cite can be accurately described as receiving results from another party.

However, the reason to focus on citations over other forms of communication is that citations do have norms which seem to be followed to some extent and, critically, they are measurable empirically as a result of the collection and systemization of academic metadata. While I'll go into the practicalities of how in the next section, citation and author data is easily available for large-scale analysis across nearly every field of study, thus this metadata has the potential to serve as a proxy for the real-world scientific communities might look like. So to create a more realistic account of network structure in science, I plan to start from this data, rather than from the idealized graphs Zollman proposed. The hope here is not that this data will prove a perfect proxy for real communication, as no single data source captures the full breadth of human communication, but as a much more empirically-motivated structure than Zollman's rings, cycles and complete graphs. However, to effectively deploy citation metadata to create realistic communication structures, I must create a reasonable rule for what constitutes evidence of communication and thus when to draw an edge.

Constructing Realistic Graphs

I divide this section into three distinct parts. First, I detail why I selected the Microsoft Academic Graph (MAG) and what specific data it contains. Then, I briefly detail my strategy for processing the large dataset and finally conclude with a more formal definition of how I define communication in terms of the specific metadata present in the MAG.

Comprehensive Metadata To infer communication from an academic social network, we first must first begin with a comprehensive dataset of academic publishing metadata. By publishing metadata, I refer to all the data associated

[4] *Publication Manual of the American Psychological Association.*
[5] Latour, "Literature."
[6] Wilhite and Fong, "Coercive Citation in Academic Publishing."

with a scholarly journal article, conference paper or book except for the actual text of the work itself. A good way to think about such metadata is everything one would put in a works cited page, so the title, a list of authors, the date of publication, the journal, etc. However, most datasets also go beyond this metadata and also include a set of citation links to other papers, a set of fields of study as assigned by publishers as well as links to other works which reference that paper.

Luckily several comprehensive sources of this metadata exist. Clarivate Analytics compiles "Web of Science" (WoS), a proprietary dataset which has vetted and comprehensive scientific metadata spanning from 1900 to the present and including over 200 million entries. However, I decided against using this dataset, which is commonly used in bibliometric papers which analyze citation metadata, due to the closed nature of the dataset which create barriers for researchers. There is also a young project, OpenCitations, which seeks to build a citation data repository that is entirely open with vetted data from publishers, however, the project is still young and coverage sparse covering fewer than half a million works.[7] For this project, I chose to use the MAG[8] which leverages Bing's web indexing service much like Google Scholar leverages Google's search infrastructure to generate a comprehensive account of academic metadata, covering over 170 million entities. However, unlike WoS, the data is available under a permissible license (the Open Data Commons Attribution License) at no cost to the researcher, which drove me to select this dataset over the others because it presents the fewest barriers to follow on work and allows me to make my results freely available.

Processing a Large Graph The specific data in the MAG is evolving constantly, however, I use a snapshot taken in October of 2018 for all my analysis. My MAG snapshot is a "multi-graph" or a graph with different node types for authors (`Author`), papers (`Paper`), fields of study (`FieldOfStudy`) and journals (`Journals`). Furthermore, there are directed edges for citations which connect papers to papers by their listed citations (`REFERENCES`), from papers to their authors (`AUTHORED_BY`), from papers to their fields of study (`IN_FIELD`), and from subfields to their parent fields (`PARENT`). These relations are formatted as a series of very large CSV files (10-100GB each) and are not easily searchable in their raw form as a result. Such large files do not fit in RAM on any affordable machine, meaning it is not feasible to use the graph in its entirety for simulations. Furthermore, understanding and visualizing the results of a simulation of such a large size would be a difficult undertaking.

To make this data useable, I decided to use the popular graph database neo4j[9] to allow for quickly query and return smaller, more manageable chunks of the overall graph. While queries over most of the properties and relations in the MAG are possible, I decided to focus on returning edges which represent probable

[7] Peroni and Shotton, "OpenCitations, an Infrastructure Organization for Open Scholarship."

[8] Sinha et al., "An Overview of Microsoft Academic Service (MAS) and Applications," @wangReviewMicrosoftAcademic2019.

[9] "Neo4j."

paths of communication between authors. To get this large dataset into neo4j, I needed to convert the relations from the raw tab-separated CSV files (TSV) to CSVs, then use the offline importer to neo4j. To do this, I built a small conversion program in the programming language Rust (`mag-csv` in source code) which could perform the conversions quickly and neo4j's offline bulk import tool (`neo4j-import`) as importing through queries is far to slow to be feasible for a dataset of this size (the "fast" import still took several days on a laptop with a relatively fast solid-state storage drive).

The `AuthorCites` Relation Once the data was in queryable form, I turned my focus to determining the precise relations I would use to capture likely communication relations between authors. To do this, I wanted to leverage citations because they are instances where we have clear evidence in the MAG that one work has influenced another. However, citations relate papers, not authors, so the cited relation must be lifted to apply to authors. Consider two authors A_1 and A_2. We say that A_1 cites A_2 if and only if there exist papers P_1 and P_2 such that A_1 authored P_1, P_1 cites P_2 and A_2 authored P_2. Formally:

$$\begin{aligned}\texttt{AuthorCites}(A_1, A_2) = \exists_{P_1, P_2 \in \texttt{Papers}} (&\texttt{AUTHORED\_BY}(P_1, A_1) \wedge \\ &\texttt{REFERENCES}(P_1, P_2) \wedge \\ &\texttt{AUTHORED\_BY}(P_2, A_2))\end{aligned}$$

All that is required is that a single pair of papers P_1 and P_2 exist to maintain the `AuthorCites` relation. From this relation, we infer that results were communicated from A_2 to A_1 by way of A_1 reading or reacting to A_2's work, which indicates both a general awareness of the cited author and the results presented in the paper. This may be a cursory awareness if the citing author glanced at the cited paper before using the citation strategically, however, that is still transmitted information which could alter the direction of the author's research. This is no perfect relation, again, lots of types of communication happen without a subsequent citation and citations might represent little to no information transferred in some cases. However, I argue that this method does a much better job of approximating scientific community structure than a cycle, wheel, complete or any other idealized graph would by virtue of being derived from real data about real interactions.

An important point of comparison here is the co-authorship relation which is often used in bibliometric community analysis. This simply relates two authors when they have co-authored a paper together. While co-authorship is clearly a strong signal that information has been transmits, I argue it is overly restrictive to fit well with Zollman's definition of edges representing communication of results between two parties. Furthermore, when there is a sole author co-authorship would not reflect any other community members that that sole author was influenced by and this is feels wrong especially when the author cited other papers. Thus, I chose to go with the `AuthorCites` relation.

While the `AuthorCites` relation does capture what we want at a high level, it does not help limit the overall data processed in a given query. To do this, I create the `AuthorCitesInField` which limits connections to only those in which P_1 and P_2 are both in the field of study of interest, $F \subseteq$ Papers. Thus, the definition becomes:

$$\begin{aligned}\texttt{AuthorCites}(A_1, A_2) = \exists_{P_1, P_2 \in F}(&\texttt{AUTHORED\_BY}(P_1, A_1) \wedge \\ &\texttt{REFERENCES}(P_1, P_2) \wedge \\ &\texttt{AUTHORED\_BY}(P_2, A_2))\end{aligned}$$

Because we are focused on specific communities, this definition works well while significantly reducing query complexity by only quantifying over all papers in a single field, not all papers in the MAG. This does ignore interdisciplinary work, which is an unfortunate downside to this approach, however, the decrease in query complexity achieved here is what makes these queries feasible at all using a relatively small machine.

Given the definition of `AuthorCitesInField`, I crafted the following neo4j query. The query takes on a different form than the definition to better leverage the graph structure and relations that exist in the graph. There is no relation which links authors to fields of study, so it is much faster in practice to enumerate a list of papers in a field than a list of authors. This query, as written, will result in duplicate relationships between authors, however those are more effectively deduplicated in a graph processing library rather than in the database.

```
MATCH (p1:Paper)-[:IN_FIELD]->(parent:FieldsOfStudy{id: $id})
WHERE p1.citationCount > 0
MATCH (p1:Paper)-[r:REFERENCES]->(p2:Paper)
MATCH (p2:Paper)-[:IN_FIELD]->(parent:FieldsOfStudy{id: $id})
WHERE p2 <> p1 AND p2.citationCount > 0
MATCH (p2:Paper)-[:AUTHORED_BY]->(a2:Author)
MATCH (p1:Paper)-[:AUTHORED_BY]->(a1:Author)
WHERE a1 <> a2
RETURN a1, a2
```

Selecting Subgraphs

Now that I've described my method for relating authors via citations, I tried the method out on several areas of research. To chose a diverse set of fields of research, I used the MAG's fields of study as groupings and its hierarchical organization of them. This organization places each field in a tree which descends from a top-level field which has no parent.

I started by selecting both "social epistemology", because that is Zollman's field of Philosophy; and "peptic ulcer" because that is the field of medicine Zollman

focuses on in his case study. After picking these, I picked "brain morphometry" as a subfield of neuroscience, "monetary policy" as a subfield of economics, "abstract algebra" as a subfield of math and "phonetics" as a subfield of linguistics. My goal with this selection of fields is to find fields of different sizes and with likely somewhat different citation practices and cultures to ensure that my tests of Zollman's model do not result in field-specific results.

In Tables 1 and 2, I detail the resulting `AuthorCitesInField` graphs in relation to co-authorship graphs, which serve as a point of comparison. I report node and edge counts to give a general idea of the size of a field in Table 1, where nodes are authors and edges are `AuthorCitesInField` relations which connect authors to author's they've cited via directed edges. Thus, the more nodes, the bigger the field and the more edges, the more well connected the field. I report basic stats pertaining to the connected components (both strong and weak) in Table 2. This includes both a count of the total number of separate components as well as the size of the largest one. It is important to note that because co-authorship is undirected, strong and weak connectivity are the same for those networks. Connectivity is a crude measure of the cohesion of a community which partition the graph into sections such that no edge goes between sections. These partitions of the graph represent fractures in the community such that authors in separated components have never cited anyone in any of the other sections. Having many large components suggests either that fields are themselves fractured or that the MAG mislabeled some author's work, leading them to be an island in a spuriously assigned field. For larger fields, it makes sense that the field would be partitioned as it seems unlikely that

A striking characteristic of most of the `AuthorCitesInField` and co-authorship fields is the emergence of singular large weakly connected components which contain most of the nodes in many, but not all, of the selected fields. However, in nearly all cases, the weakly connected component is much larger than the strongly connected component. This effect is commonly observed in large graphs and most famously first observed in the analysis of the early web,[10] where the large weakly connected component did not imply a large weakly connected component. These results emphasize the structural differences between directed and undirected graphs, in that simply allowing the `AuthorCitesInField` to be bidirectional can unify a fractured field. Because Zollman's models are only meant to be run on connected graphs, the model will, in practice, be run on connected subgraphs rather than the overall fields.

Table 1: Selected fields and resulting `AuthorCitesInField` networks and Co-Authorship and associated node and edge counts.

network	nodes	edges
social epistemology AuthorCited	651	1490

[10] Broder et al., "Graph Structure in the Web."

network	nodes	edges
social epistemology CoAuthor	1663	3777
brain morphometry AuthorCited	5154	77371
brain morphometry CoAuthor	8584	139059
monetary policy AuthorCited	17620	454030
monetary policy CoAuthor	30418	108882
abstract algebra AuthorCited	311	968
abstract algebra CoAuthor	977	2400
peptic ulcer AuthorCited	20970	562167
peptic ulcer CoAuthor	49017	317141
phonetics AuthorCited	5766	58192
phonetics CoAuthor	10399	36707

Table 2: Selected fields and resulting `AuthorCitesInField` (AC) networks and Co-Authorship (CA) and associated strongly connected component (SCC) counts, weakly connected component (WCC) counts, size of largest strongly connected component and size of largest weakly connected component.

network	largest SCC	num SCCs	largest WCC	num WCCs
social epistemology AC	10	615	560	30
social epistemology CA	82	596	82	596
brain morphometry AC	1661	3392	5049	11
brain morphometry CA	5186	576	5186	576
monetary policy AC	9616	7913	17541	32
monetary policy CA	13306	5974	13306	5974
abstract algebra AC	12	285	192	29
abstract algebra CA	25	337	25	337
peptic ulcer AC	8383	12444	20449	67
peptic ulcer CA	14698	7371	14698	7371
phonetics AC	2049	3640	5608	29
phonetics CA	3734	2064	3734	2064

Part 2: Verifying Zollman's Model

Replication of Zollman's model

Zollman's Model on Empirical Graphs

New Measures

Conclusion

Broder, Andrei, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar

Rajagopalan, Raymie Stata, Andrew Tomkins, and Janet Wiener. "Graph Structure in the Web." *Computer Networks* 33, no. 1 (June 2000): 309–20. https://doi.org/10.1016/S1389-1286(00)00083-9.

Collins, Kimberley, David Shiffman, and Jenny Rock. "How Are Scientists Using Social Media in the Workplace?" *PLOS ONE* 11, no. 10 (October 2016): e0162680. https://doi.org/10.1371/journal.pone.0162680.

Latour, Bruno. "Literature." In *Science in Action : How to Follow Scientists and Engineers Through Society*, 21–62. Cambridge, Mass. : Harvard University Press, 1987.

"Neo4j," n.d.

Peroni, Silvio, and David Shotton. "OpenCitations, an Infrastructure Organization for Open Scholarship." *Quantitative Science Studies* 1, no. 1 (January 2020): 428–44. https://doi.org/10.1162/qss_a_00023.

Publication Manual of the American Psychological Association. 6th ed. Washington, DC : American Psychological Association, 2010.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, and Kuansan Wang. "An Overview of Microsoft Academic Service (MAS) and Applications," May 2015.

Wang, Kuansan, Zhihong Shen, Chiyuan Huang, Chieh-Han Wu, Darrin Eide, Yuxiao Dong, Junjie Qian, Anshul Kanakia, Alvin Chen, and Richard Rogahn. "A Review of Microsoft Academic Services for Science of Science Studies." *Frontiers in Big Data* 2 (December 2019): 45. https://doi.org/10.3389/fdata.2019.00045.

Wilhite, Allen W., and Eric A. Fong. "Coercive Citation in Academic Publishing." *Science* 335, no. 6068 (February 2012): 542–43. https://doi.org/10.1126/science.1212540.

Zollman, Kevin J. S. "The Epistemic Benefit of Transient Diversity." *Erkenntnis* 72, no. 1 (October 2009): 17. https://doi.org/10.1007/s10670-009-9194-6.