Proposing Ties in a Dense Hypergraph of Academics

Aaron Gerow[1*], Bowen Lou[12], Eamon Duede[1], and James Evans[13]

[1] Computation Institute, University of Chicago, Chicago, USA
[2] Wharton School of the University of Pennsylvania, Philadelphia, USA
[3] Department of Sociology, University of Chicago, Chicago, USA
* Corresponding author: gerow@uchicago.edu

Abstract. Nearly all personal relationships exhibit a multiplexity where people relate to one another in many different ways. Using a set of faculty CVs from multiple research institutions, we mined a hypergraph of researchers connected by co-occurring named entities (people, places and organizations). This results in an edge-sparse, link-dense structure with weighted connections that accurately encodes faculty department structure. We introduce a novel model that generates dyadic proposals of how well two nodes should be connected based on both the mass and *distributional similarity* of links through shared neighbors. Similar link prediction tasks have been primarily explored in unipartite settings, but for hypergraphs where hyper-edges out-number nodes 25-to-1, accounting for link similarity is crucial. Our model is tested by using its proposals to recover link strengths from four systematically lesioned versions of the graph. The model is also compared to other link prediction methods in a static setting. Our results show the model is able to recover a majority of link mass in various settings and that it out-performs other link prediction methods. Overall, the results support the descriptive fidelity of our text-mined, named entity hypergraph of multi-faceted relationships and underscore the importance of link similarity in analyzing link-dense multiplexitous relationships.

1 Introduction

High impact research commonly spans fields of science, leading universities to increasingly focus on ways to catalyze cross-disciplinary collaborations. Some institutions have sought to overcome challenges of disciplinary compartmentalization by implementing research networking and researcher profiling systems[4]. However, there is little evidence to suggest that such systems generate new, effective collaborations that span traditional boundaries [32]. Inter-disciplinary institutes may be positioned to play the match-maker, but these are also difficult to systematically qualify [15]. An important challenge, then, is to represent relationships among academics from which novel, productive links can be reliably modeled. The challenge is compounded by the fact that actual collaborations are a sparse structure: most people do not collaborate with most others. This paper

[4] profiles.catalyst.harvard.edu is one example.

proposes a method of mining a hypergraph with a large number of hyper-edges — 25 times the number of nodes — and a model which uses an edge-based, distributional measure of link similarity to propose relationships among academics.

Network mining and modeling bears on a range of problems in computer systems, power distribution, cell-biology, cognition, organizational structure and even terrorist networks [3, 12, 33]. Mining social structure, particularly from text, is an important task for automated recommendation systems [18]. Perhaps the simplest form of social recommendation is proposing ties between people in a social network. This task, known generally as link prediction, can lend insights to how networks grow and change over time, and is the subject of a great deal of research [28–30, 42]. However, link prediction has been almost exclusively explored in the context of single-mode, unipartite networks where edges only connect two nodes [5], which over-simplifies the multiplexity of social dynamics. In reality, nearly all personal relationships consist of multiple, sometimes diverse connections that vary in strength, breadth and meaning. Here, we show that such multiplexity can be represented using *named entities* (NEs) in faculty CVs. These entities can be interpreted as the relationships in which people participate [13], and can consist of organizations (publishing in the same journal, sharing professional affiliations or committee membership), location (cohabiting an office, building or city) and personal relationships (co-authors, supervisors or mentors). Connecting academics by the co-occurring NEs observed in their CVs forms a complex hypergraph that is edge-sparse (most edges connect only few nodes) and link-dense (most dyads are connected by at least one edge)[5].

The social recommendation task in an edge-sparse, link-dense hypergraph is similar to traditional link prediction given its focus on dyadic links. The goal is to propose ties (in our case, weighted links) between nodes based on the structure of the network. Because most nodes are already connected in a link-dense graph, recommendations can be interpreted as how strong a link *should be* based on a dyad's local structure. As we will see, this density is problematic for link prediction methods that rely on variation in the neighborhood around a dyad. To overcome this density, the proposals should account for both the *strength and similarity* of links through a dyad's neighbors. Fortunately, we can use the density to our advantage by treating edges as points in a space defined by nodes' edge-incidence vectors, allowing similarity to be measured as inverse distance in high-dimensional space. This strategy is particularly advantageous as it allows similarity to be calculated without using external information.

There is likely some interplay between the mass and similarity of links on either side of a neighbor. We hypothesize that the mass of a link through a transit node is only "available" to the extent links are similar. That is, similar links can use most of the edge mass, while dissimilar links use only a fraction. In figure 1b, this is analogous to saying that the mass of link **AB** (edges 1, 2 and 3) plus **BC** (edges 1, 4 and 5) is proportional to the similarity of **AB** and **BC**. We hold that this intuition is not only socially plausible, but that it can be used to more accurately recommend connections.

[5] We use *link* to refer to the set of hyper-edges connecting a pair of nodes.

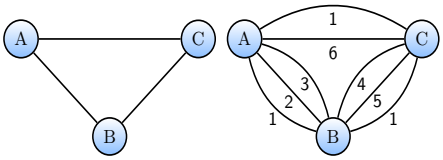

Fig. 1. A triad of nodes A, B and C as a simple network (a; left) and as a hypergraph (b; right). In (b), note that not only can nodes be connected by more than one edge (ie. a multi-mode graph), but edges can connect more than two nodes as with edge 1. We write **AB** to denote the link from A to B, here consisting of edges $\{1, 2, 3\}$.

After reviewing related work and providing some background in the next section, section 3 introduces our model, including definitions of link mass and distributional link similarity on weighted hypergraphs. Section 4.1 presents four experiments where the model is used to recover link mass from four versions of the data, lesioned in different ways. Section 4.2 compares our model to other link prediction methods after which we review the results of the evaluations. The final section discusses the contribution of the representation and associated model, pointing to future work and applications.

2 Background & Related Work

Network analysis is a widely studied and increasingly important area of research in biology, economics, natural language processing, sociology and computer science [12, 16, 17, 37]. In particular, social network analysis seeks to model how people (nodes) develop relationships to one-another (edges). Such social organization tends to form groups where similar kinds of ties are shared among the members [10, 11]. To represent social interactions in a simple network, the meaning of edges must be globally defined, perhaps denoting friendship, kinship or shared affiliation. A hypergraph, on the other hand, can represent relations that connect any number of people, who in turn may have any number of connections [4, 31, 37]. Figure 1 depicts a simple network and a hypergraph with six hyper-edges, where each dyadic link consists of a set of hyper-edges. What these edges are, that is to say their semantics, may vary from one graph to another, but as we will see, people, places and organizations offer a reasonable starting-point for defining academic relationships.

Scott Feld [13] suggested that social circles, such as affiliations, beliefs and practices, make up focal points around which people engage and connect. These foci naturally comprise a network structure that interrelates people as nodes connected by the foci in which they participate. The current study operationalizes Feld's *focus theory* as a model on hypergraphs where people are nodes and foci are edges. Our approach effectively extends focus theory to a high-dimensional space where the modes of interaction (edge-types) can greatly out-number the actors themselves. It is important to note that foci need not be restricted to social circles or affiliations — they may also be people themselves. Using people,

places and organizations as edges will allow our analysis to retain the intuitions from social network analysis and enable us to represent complex relationships [22]. The challenge will be to propose new, or at least stronger ties in a structure that is already rich with connections.

Recently, some work has explored statistical learning methods on networks to predict academic relationships [21], which combines advances in machine learning with social network analysis in a unipartite setting. Similar to the work here, given the use of NEs as features in the network, is a method of extracting names from websites to mine communities of people [27]. Some work on layered social networks implicitly adopts the hypergraph representation [25, 26, 44], though most social network analysis still relies on single-mode, unipartite graphs. What's more, most research using hypergraphs is framed as n-partite graph analysis, typically with a small number of layers [4, 24]. Here, we explore hypergraphs where the layers far out-number the nodes themselves. Although this certainly compounds the representational complexity of a network, it is considerably more plausible as it allows actors to relate in the myriad ways people do in the real world. Hypergraphs have been used to predict multi-actor collaborations (ie. teams) in academic social networks [36], by predicting the formation of links via new and old hyper-edges. Because academic teams tend to evince relationships in different and systematic ways, the hypergraph is a particularly robust data-structure for modeling team formation [40], recommending new collaborations [43] and as a model of how scientists actually seek out new projects [37]. In this paper, instead of assuming a set of edge-types, we show that a distributional measure of link similarity can help leverage the abundance of edge-types found in a text-mined hypergraph of NEs.

Proposals in our model are effectively an estimate of how connected two nodes *should* be, and can be thought of as link predictions in a static setting. Traditional link prediction is the task of assigning a likelihood of observing new a connection at some future time using information about the current network (see [28] for a review). Getoor et al. [19] introduce a framework for representing and learning probabilistic link structure in arbitrarily complex relational data. Extending the statistical learning approach, Al Hasan et al. [2] develop a supervised learning method that uses structural and "aggregate" features to train a classifier to predict links. Although there are some hybrids, link prediction methods tend either to allow the use of global information about the network (diameter, path-distances, etc.) or limit themselves to local information (degree, shared neighbors, etc.). The method in this paper is more similar to the latter, as it restricts the search space for dyadic proposals to the links through a dyad's shared neighbors.

We hypothesize that links are characterized by two related qualities: their edge-mass and their similarity to other links. Mass accounts for the strength of a link, regardless of its composition. The weights of individual hyper-edges can then be used to discriminate, for example, between participation in a large conference where two people are unlikely to meet, and more intimate relationships, like a workshop or committee.

3 Method

Our goal is to propose ties in a social network that is already quite dense with links. Two nodes are considered linked if they are connected by one or more hyper-edges. A link refers to a dyadic connection whereas a hyper-edge can connect any number of nodes (see figure 1). Proposals, then, can be calculated for any pair of nodes, connected or not, referred to as a dyad.

Proposals are calculated using transitivities, with the assumption that shared neighbors provide a strong indication of how actual people tend to introduce other people to one-another [6, 35]. Although restricting the search space to shared neighbors can be problematic, especially in sparse graphs where higher-order structure is important [39], it is both plausible to assume academics limit their search behavior in this manner and it is known to successfully characterize social networks more generally [38]. Also, this formulation allows users of the model to recover straight-forward justifications for individual proposals — something discussed in the final section. Specifically, a proposal for a pair of nodes, n_1 and n_2 in a hypergraph, G, is calculated as follows:

$$proposal(n_1, n_2; G) = \sum_{t \in \Gamma(n_1) \cap \Gamma(n_2)} S(\mathbf{n_1 t}, \mathbf{n_2 t}) \tag{1}$$

where $\Gamma(x)$ denotes the neighbors of node x and $\mathbf{n_x t}$ is the set (hence in bold) of edges connecting nodes n_x and t. S is a function relating two sets of weighted hyper-edges to be defined below.

3.1 Link Mass

To account for the strength of participation in a given edge, for example membership in the Association for Computing Machinery (ACM), an edge's node-incidence is weighted by the number times it was mentioned in someone's CV. This helps account for the difference between someone who mentioned the ACM 100 times and someone who mentioned it only once. There is no limit to the frequency of an edge's weight per-node, nor is there a limit to the total number of entities by which a node may be connected. However, to accommodate the inverse relationship between the number of participants in an edge and its importance, we normalize the weights of each edge by their sum. Formally, we take the edge-wise L1-norm of G, which makes edge-weights inversely proportional to the number of nodes they connect; an edge that only connects two nodes, receives a weight of .5, whereas edges that connect many nodes receive a weight closer to zero. Although this normalization down-weights predominant edges, because entities are initially coded as occurrence frequencies, there is still room for variation in a how much nodes participate in even the most frequent edges. Figure 2 shows an example of the initial, frequency-weighted matrix and its normalized counterpart.

One simple way to think about the quality of a link is by its mass. Socially, this can be thought of as the sum of all the ways, important and trivial, two people relate. We assume that edges in a graph, G, have weights associated individually with the nodes they connect. That is, each node may "participate" in an edge to a varying degree (ranging from 0 to 1). The weight for a dyadic link, then, is the sum of these edge-incidence weights. The weight for a link between two nodes *through a transit node* is the minimum of the two links' mass, as this represents the maximal relation two people have through either link. This minimum is the weight used to compute dyadic link mass from one node to another, though an intermediate node, t:

$$Mass(n_1, t, n_2) = min(\sum_{i \in \mathbf{n_1 t}} w_i, \sum_{i \in \mathbf{n_2 t}} w_i) \tag{2}$$

where w is the edge's weight as described above. Link mass only accounts for the combined strength of edges through a transit node. Referring to figure 1, link mass from A to C through B would be the sum of the edge-weights in **AB** and **BC**. While this may be a plausible way to model how people recommend relationships to one another (by the magnitude of their relationship to each person), it fails to account for the edges themselves.

3.2 Link Similarity

Link similarity is an important component when examining the social closure of transitivities: two people are more likely to be introduced if they have *similar* links through a shared acquaintance than if those links are dissimilar. To address link similarity, we conceive of edges as points in the normalized space defined by edges' node-incidence. A single edge is represented as the coordinate given by its weighted node-incidence vector in G, which in the data for our experiments is highly dimensional. The similarity of two links, $\mathbf{n_1 t}$ and $\mathbf{n_2 t}$ (ie. the connection from n_1 to n_2 through t), is defined as one minus the Euclidean distance between each edge-sets' centroid, $C_{\mathbf{n_1 t}}$ and $C_{\mathbf{n_2 t}}$:

$$\begin{aligned} C^{(d)}_{\mathbf{n_1 t}} &= \frac{1}{|\mathbf{n_1 t}|} \sum_{i \in \mathbf{n_1 t}} w_i \in [0,1]^d \\ C^{(d)}_{\mathbf{n_2 t}} &= \frac{1}{|\mathbf{n_2 t}|} \sum_{i \in \mathbf{n_2 t}} w_i \in [0,1]^d \\ Sim^{(d)}(n_1, t, n_2) &= 1 - \sqrt{(C^{(d)}_{\mathbf{n_1 t}} - C^{(d)}_{\mathbf{n_2 t}})^2} \end{aligned} \tag{3}$$

where d is the dimensionality of the normalized node-space. Using nodes to define a normalized space in which to compare edges allows a completely internal, data-driven conception of similarity. This feature of our model not only frees it from potential problems with extrinsically defined similarity metrics, such as under-specification, incompleteness or incorrectness, it also scales well to graphs with arbitrarily large number of edge-types. While there is no reason an externally

defined similarity measure could not be used in its place, the distributional measure allows our proposals to be derived within-model.

With these definitions of link mass and link similarity, we can test our hypothesis that the strength of links through a transit are in fact dependent on their similarity. Further, because we assumed the node-incidence weights used to define each edge are normalized from 0 to 1, we can use them to scale the link mass by multiplying the values of Sim by $Mass$. The intuition is that greater participation in similar, intimate venues increases the likelihood two people connect. Substituting this combination of Sim and $Mass$ for S in the first equation, we can define the strength of a proposal between two nodes, n_1 and n_2:

$$proposal(n_1, n_2; G) = \sum_{t \in \Gamma(n_1) \cap \Gamma(n_2)} Mass(n_1, t, n_2) * Sim(n_1, t, n_2) \quad (4)$$

After generating proposals, p, the values are normalized: $p_i \leftarrow \frac{p_i - p_{min}}{p_{max} - p_{min}}$.

4 Results

To evaluate the method outlined above, we retrieved a set of 2,511 CVs of faculty members at several large, research-intensive universities[6]. Each CV was coded by department[7]. The Stanford Named Entity Recognizer (NER) was used to extract NEs from the texts of each CV [14, 41]. The Stanford NER extracts people, locations and organizations with a reported F_1-score of 0.93 (precision = 0.93, recall = 0.92). These entities define the hyper-edges in the network. Note this means an academic may be a node *and* a hyper-edge. This process yielded 802,131 unique hyper-edges (27,982 locations, 142,350 persons, 230,739 organizations). All NEs that connected at least two people in the network were retained. This has the side-effect of filtering out spurious entities that were only observed once. Though most edges do not connect most nodes, only 14 nodes were not part of the largest connected component, to which we restrict our analysis. In fact, the hypergraph was almost fully connected: most nodes were connected to all other nodes by at least one edge, which is not surprising given that each university constitutes an ORGANIZATION edge connecting a large number of academics. In total, there were 3,116,256 dyads of which 66% were directly connected by an average of 2.2 hyper-edges. This hypergraph comprises the $N \times E$ matrix referred to as G, an example of which is depicted in figure 2. After link-mining, the matrix was weighted and normalized by the method described above. A preliminary evaluation of this representation was performed where the node-incidence vectors $\{E \in G\}$ were used as features to classify academics in their departments. We found that, when restricted to the same number of features, token-based feature-sets are significantly outperformed by our representation (see the appendix).

[6] Available at klab.ci.uchicago.edu/data/CV_data.tar.gz

[7] 122 faculty members had appointments in more than one department.

	Observed					Normalized				
	n_1	n_2	n_3	n_4	n_5	n_1	n_2	n_3	n_4	n_5
W3C	1	2	0	5	2	0.1	0.2	0.0	0.5	0.2
ACM	1	0	0	3	0	0.25	0.0	0.0	0.75	0.0
IEEE	0	0	2	0	0	0.0	0.0	1.0	0.0	0.0
SIGKDD	0	0	1	1	0	0.0	0.0	0.5	0.5	0.0

Fig. 2. Example $N \times E$ matrix defining G. Nodes / CVs are represented in each column with four example NEs / edges as rows. Shown are the initial matrix (left), where the numbers represent an entity's frequency in each CV, and the normalized matrix (right).

4.1 Recovering Link Mass

The goal of the proposal model described in section 2 is to generate recommendations of stronger ties. Because actual relationships are formed over the course of time, proposing ties in networks is often framed as predicting whether or not a link will form in the future [28]. In the absence of temporal data, we evaluated our model in a static setting. We describe its performance in four settings where the observed graph, G, was systematically lesioned, after which proposals were calculated on the resulting graph, G'. In each experiment, the goal is for proposals generated on G' to replicate or "recover" the link mass observed in G. Because we make no assumptions about the distribution of link-masses in G, we use Spearman's rank-correlation to evaluate whether the *ordering* of dyadic proposals on G' is similar to that observed in G. By computing this correlation at rank from the strongest to weakest links in G, we can also test whether our proposals tend to be more or less accurate as a function of the initial link strength. These experiments exhaust a reasonable space of evaluation, showing that our model is robust to local and global lesioning of both links and hyper-edges (figure 4). In all cases, we compare the model to a random model.

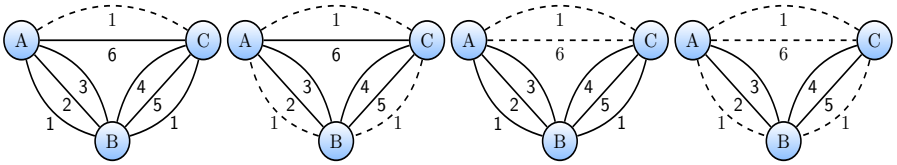

Fig. 3. Four versions of G' corresponding to four evaluation schemes, derived from the example triad in figure 1b. Dashed lines represent deleted hyper-edges that were initially present in G. From left to right: (a): Local edge deletion, where one component of a hyper-edge in a dyad is deleted. (b): Global edge deletion, where a sample of hyper-edges are deleted from the entire graph (here hyper-edge 1 is deleted). (c): Local link deletion, where all edges connecting a dyad are deleted. (d): Global link deletion, where the edges comprising the chosen dyadic link are deleted from the entire graph (here, because link $\mathbf{AC} = \{1, 6\}$, hyper-edges 1 and 6 are deleted globally). In all evaluations, dyadic proposals are generated from nodes analogous to $\mathbf{AC} \in G'$ over all shared neighbors (those in a similar position to B).

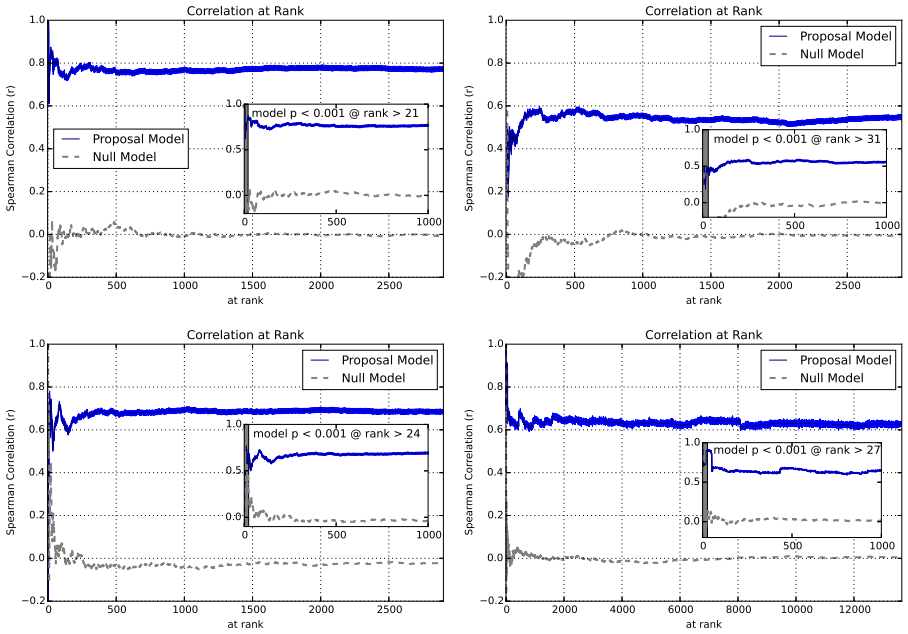

Fig. 4. Correlation at rank for the four link-mass recovery experiments in section 4.2. Results for each lesioning scheme (figure 4) clock-wise from top left: local hyper-edge deletion, global hyper-edge deletion, local link-deletion and global, like-selected hyper-edge deletion. In each, rank refers to the inclusion of the n strongest dyadic links in G (equivalent to **AC** in figures 1 and 3) of which in-lays show the top 1,000. The shaded region in all plots denotes +/- 1 SD of the mean over 10 folds with Fisher's z-transformation. In all experiments the model results (solid blue line) are significantly better than a null model (dashed grey line).

Local Hyper-Edge Deletion. The first experiment of the recommendation model tests how well the proposal scores can compensate for locally deleted hyper-edges. To simulate a world where two people in our network failed to relate in a particular way, a random hyper-edge is deleted from a dyad, for which a proposal is then generated using the rest of the graph. In figure 4a, this is analogous to removing hyper-edge 1, locally from the **AC** link. The proposal for $A \rightarrow C$ should compensate for the loss of the mass of hyper-edge 1. The proposals correlated positively with a mean Spearman $r = 0.77$ ($p < 0.0001$; mean of 10, 10% samples; using Fisher's z-transformation) for all dyads. If we use only transitive link mass to generate proposals (without link similarity) the correlation is positive but less strong at $r = 0.69$ ($p < 0.0001$): the use of link similarity improved the model by about 12%. Additionally, there was no discernible trend in the correlation at rank (figure 5a) from the strongest to weakest links in the original graph, G, implying the model performs comparably well for strong and weak links alike. The model also out-performs a null model that draws a random dyadic mass from G, showing the results are not an artifact of the mass distribution.

Global Hyper-Edge Deletion. In the second experiment, the model was used to generate proposals on a globally, edge-lesioned version of G. In this task, 20% of the hyper-edges in the observed network were removed, and proposals were generated on the remainder. Even after this lesioning, most dyads were linked by at least one hyper-edge. Figure 4b shows a triad in which hyper-edge 1 has been removed, which affects each dyad because it connects all three nodes. Each run of the model on this evaluation generates a proposal for every dyad, but the edge-removal process selects a random 20% of hyper-edges, therefore, a 10-fold cross-validation scheme was employed. The model was judged to be successful to the extent that the proposals drawn on the lesioned graph, G' correlate to the original link masses in G. On average, the proposals were found to correlate with original link masses at a Spearman $r = 0.55$ ($p < 0.0001$, 10 folds). With proposals that use link mass alone, the correlation is still positive, but less strong at $r = 0.49$ ($p < 0.0001$). This indicates the link similarity is approximately as important on this task as it was in the first evaluation. Figure 5b shows the correlation at rank. The lack of trend in this statistic implies the model did as well for strong links as for weak. The model also out-performs the null model.

Local Link Deletion. This evaluation tests the ability of the model to recover the mass of entire links. Here, a dyad is completely disconnected, for which a proposal is then generated. In Figure 4c, this is equivalent to removing hyper-edges 1 and 2. The strength of the generated proposals, then, should correlate to the original dyadic link masses. On the local link deletion task, the proposals correlated to the original dyads with a Spearman $r = 0.62$ ($p < 0.0001$; 10, 10% folds). Using mass alone produced a slightly weaker correlation of $r = 0.61$ ($p < 0.0001$) which shows that on this task, link similarity is less important than on the previous two evaluations, providing less than 1% performance gain. Figure 5c shows the correlation at rank. Overall, the model performs significantly better than the null model, reiterating that the findings are not an artifact of the distribution of link masses in G.

Global, Link-Selected Hyper-Edge Deletion. The final experiment tests the ability to compensate for hyper-edges selected from links and deleted globally. For randomly selected dyads in G, the hyper-edges comprising the link are removed from the entire graph. In figure 4d, nodes A and C are connected by hyper-edges 1 and 6, which are then deleted throughout the network to yield the lesioned G'. Proposals were generated on G' and their correlation to the the original dyadic masses is reported. This experiment is effectively a combination of experiments 2 and 3: the model must provide proposals that compensate for an entirely missing link (experiment 3) using a graph with fewer edges than initially observed (experiment 2). The proposals in this configuration correlated positively with the link masses in G: Spearman $r = 0.69$ ($p < 0.0001$; 10, 10% folds). This was stronger than when using link mass alone, which yielded $r = 0.57$ ($p < 0.0001$). Figure 5d shows the correlation at rank for this experiment, in which there is no discernible trend: the model is consistent across strong and weak links in G.

4.2 Proposals as Predictions

The preceding experiments show that our model can recover the dyadic link mass in G after it has been systematically lesioned in different ways. This section tests our model as a method of link prediction. Link prediction is typically thought of as the task of predicting a link at some future time given historical data. Because our mined hypergraph lacks temporal data, we employ a static variant of link prediction. As in other formulations, a link refers to a single dyadic connection. Although some work has explored predicting hyper-edges that connect more than two nodes [21, 40], in our model, though it uses hyper-edges to define link similarity, it generates dyadic proposals more analogous to links in a unipartite graph. To compare our proposals to other link prediction methods, we simplify G into a unipartite graph, $\widetilde{G}$, by treating links (sets of edges in G) as unique, unlabeled edges. Additionally, because our proposals are not simply yes or no binary suggestions as with typical link-prediction tasks, evaluations like precision, recall and ROC / AUC are less well-suited to assessing the proposal scores. As such, we again use Spearman rank-order correlation to compare predictions that preserve the rank of link-strengths observed in G.

We compare our model to three other link prediction methods: the Jaccard coefficient, preferential attachment and the Adamic-Adar method [1, 28]. The Jaccard coefficient is a structural measure defined as the size of the intersection of a dyads' common neighbors over that of their union. Preferential attachment metricates the intuition that highly connected nodes are more likely to form links than less connected nodes. Adamic-Adar predictions account for the size of shared neighbors' neighbors, defined as the inverse log of the number of every shared neighbor's neighbors. Each method is defined as follows:

$$\text{Jaccard Coefficient:} \qquad \frac{|\Gamma(n_1) \cap \Gamma(n_2)|}{|\Gamma(n_1) \cup \Gamma(n_2)|}$$

$$\text{Preferential Attachment:} \qquad |\Gamma(n_1)| \cdot |\Gamma(n_2)|$$

$$\text{Adamic-Adar:} \sum_{x \in \Gamma(n_1) \cap \Gamma(n_2)} \frac{1}{\log |\Gamma(x)|}$$

We compared our model to the three link prediction methods described above using variants of the link-centric experiments in the preceding section (figure 3c and d). In the first link prediction task, we remove a random 10% of the links in G, to yield G', on which which our model is run. For the other link prediction methods, G' is converted to a unipartite graph, $\widetilde{G}'$ from which predictions are generated. Because $\widetilde{G}'$ includes even the weakest links in G', it is quite dense. We varied the threshold required for a link to be considered and report performance on resulting graph. That is, the weakest dyadic ties (the same links in G' as in $\widetilde{G}'$) were successively removed over 40 cuts until the graph was empty. Figure 5a shows the results of each method under this local, link-deletion scheme. A high correlation implies that a method's predictions preserved the ranking of links in the initial graph, which is a stronger indication of performance than a binary distinction for extant links.

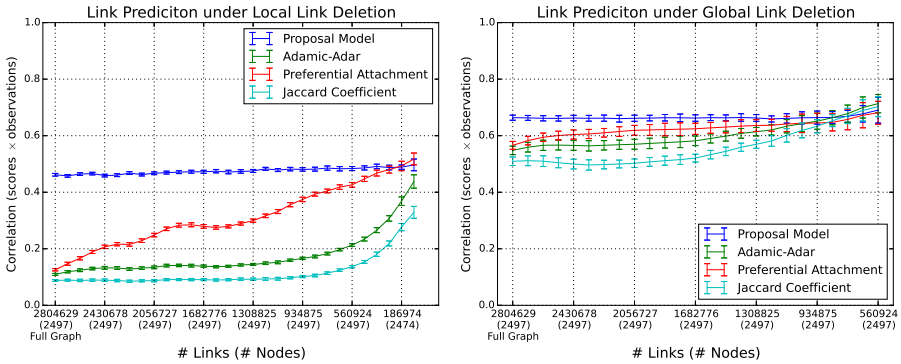

Fig. 5. (a; left): Spearman rank-correlation of predictions vs. observations on a randomly removed selection of 10% of the links. The lesioned graph, G' for the proposal model and $\widetilde{G}'$ for the other methods, was used to generate predictions. As weak links are successively removed from consideration (horizontal axis), the graph becomes increasingly sparse. After 36 cuts, the graph fractured into separate components. **(b; right)**: Analogous correlation on a random selection of 10% available links having been removed with their comprising hyper-edges (figure 4d). Here, the graph fractured after 30 cuts. All error-bars are +/- 1 SE of the mean across 10 runs.

Figure 5a shows that when all links, weak and strong, are considered, the graph's density is detrimental to the performance of the unipartite methods. The performance of these methods, which do not account for link similarity, underscores its importance in dense settings where structural information is more uniform. The results also show that our model is considerably more stable with regards to the effect of weak links. The Jaccard coefficient and the Adamic-Adar method perform weakly until the number of links is reduced to the strongest 15%, after which they begin to rise. Preferential attachment exhibits similar behavior to the other models at either end of the link density spectrum, but performs steadily better as weak links are pruned. Our model is not only more consistent, its performance is only matched (by preferential attachment) when the links are reduced to their strongest 8%.

In a second link prediction experiment, the task is designed similarly to the final experiment in section 4.1. The setup is the same as the link prediction experiment above, except that when a link is selected for removal, all the hyper-edges comprising it are removed from the graph (see figure 3d). After this lesioning yields G', it is again simplified to $\widetilde{G}'$ for use with the other link prediction methods. Figure 5b shows the rank-correlations of the four methods in this task over the increasingly sparse graph. The results show that the added sparsity, due to the removal of additional hyper-edges during the lesioning process, gives all methods a boost. The proposal model still out-performs the other methods until the graph consists of only the strongest links.

4.3 Discussion of Results

Altogether, our model performs well across a range of evaluations. However, we make no claim that the algorithm is generalizable to hypergraphs of arbitrary structure, but for the purpose of recommending ties in an already densely connected social network, it is quite successful. Some of the model's strength is due to the data-structure itself, which is not only a novel operationalization of Feld's focus theory of social interaction [13], but also accurately realizes departmental structure (see the appendix). The four experiments in section 4.2 show that in many conceivable situations, corresponding to the respective lesioning schemes, the proposals can account for situations where connections either were not properly mined / observed, or have actually not been made. On the link prediction tasks in section 4.3, our model out-performs three other methods that use local structure. This setting is slightly unfair to the traditional link prediction methods because they are unable to account for link similarity. Because the unipartite graph $\widetilde{G}$ is a simple weighted graph that is exceptionally dense, structural information like a node's degree and neighborhood are relatively unhelpful. It might be that precisely the information that is lost in simplifying G to $\widetilde{G}$ is the very information making our model more accurate. It remains to be seen whether our model performs well in a dynamic setting where historical data are used to predict future links. In a static setting, however, our results not only support our model and representation, they also point to an important social dynamic: that in multi-faceted relationships, both the magnitude *and* similarity of ties are important in developing connections.

Though the results here support the veracity of proposals from our model, they could be evaluated more qualitatively by presenting experts with recommendations with the greatest difference between proposed scores and observed mass. Though this evaluation would be expensive, it would potentially uncover dimensions of scholarly relationships not accounted for in our representation. Another area for future exploration is dimensionality reduction. Unlike a number of recent approaches to tasks like collaborative filtering and community detection, our model does not employ dimensionality reduction on the observed structure (eg. [20]). Given the cell-wise sparsity of G (99.8% zero-values), we explored using sub-spaces reduced with PCA and tSNE. On the experiments in section 4.2, we found a steady degradation of performance as target the dimensionality approached 0, which implies there is little to no additional information embedded in lower-dimensional projections. This reduction amounts to a blind compression of the edges based on their points in node-space — a space we have no reason to believe is adequately represented by its most prominent components. Alternatively, one could imagine performing such compression in a semantically principled manner, for example, by collapsing similar publication venues or geographical locations, though we leave this to future work.

5 General Discussion

People are embedded in a deeply social, highly connected, dense social structure. As representations grow to account for the multiplexity of social ties, they will naturally begin to exhibit this density — something that is surely not unique to academics. Social media, for instance, has greatly added to the available ways for people to connect, ostensibly confounding traditional social network analysis. The hypergraph structure explored in this paper is able to capture the emergent multiplicities apparent in relationships, specifically among academics. By using the mass of links through two people's shared neighbors, scaled by the similarity of the same links, scores from our model were able to correlate strongly to initial observations. This confirms our hypothesis that link mass and similarity mutually influence how likely transit nodes are to connect neighbors.

One critique of the model is that it overlooks higher-order structure, which is often exploited using spectral methods [8, 34] or exponential random graph models [23, 35]. However, we found that the average dyad shares 74% of all other nodes as neighbors:

$$\frac{1}{|\mathbf{G}|}\sum_{n_1,n_2\in\mathbf{G};n_1\neq n_2}\frac{|\Gamma(n_1)\cap\Gamma(n_2)|}{|\mathbf{G}|-2}=0.74 \tag{5}$$

This means that accounting for higher-order structure is largely unnecessary. Also, by limiting our model to first-order transitivities [38], one can recover a ranking of the shared neighbors that contributed the most to a proposal. This kind of ranking would be particularly important for a qualitative analysis that sought to justify *why* proposals were made.

The model explored in this paper is primarily a social one, where academics are placed in the space of locations, organizations and people observed in CVs. It would be possible to expand the set of edge-types that populate the hypergraph considered by our method. There is also, of course, no inherent reason to restrict analysis to CVs, especially as more robust edge-types are explored that may be found elsewhere. Other sources of relational information about academics might include research statements, grant proposals and research publications, which might encode topical and methodological connections among researchers. One could imagine recovering such information from text about research topics and their respective methods of inquiry using a mixture of human annotation, machine learning and crowd curating. Though these further sources of relational information remain unexplored, we hold that the representation and model presented here provide a strong foundation — one that is data-driven and that can account for the complex, multi-dimensional nature human of relationships.

Acknowledgments

Thanks to Alex Dunlap, Joshua Beck, Ariel Gans and Michael Hochman for help gathering, cleaning and analyzing data, as well as to Bill Shi, John Goldmsith and Birali Rusheda for advice on the model. Thanks to the SWIFT team (swift-lang.org) for help parallelizing various aspects of the model and to the Open Computing Consortium for computing resources. This work was supported by the Neubauer Collegium at the University of Chicago and by a grant from the Templeton Foundation to the Metaknowledge Research Network.

Appendix: Classifying Academics

As a preliminary task, the veracity G was tested by using it to classify academics into their known departments. The expectation is that the closer two academics are in G, the more likely it should be they share a department. In this experiment, all NEs containing forms of the word *department* were discarded. Proximity was defined as the cosine similarity between two nodes' weighted edge-incidence vectors (rows in figure 2). All pairs of academics were ordered by decreasing distance and the percent matched was calculated at every rank (precision at rank; figure 6). Though some academics hold appointments in more than one department, which makes a random guess slightly easier than $\frac{1}{departments=90}$, a null model that guesses the most frequent department (Economics, $N = 194$) provides a baseline. Note that this classification model is not learned or fit to the data, rather, it simply shows that academics with similar relationships, defined by the weighted edge-incidence vectors, tend to be in the same department. While there is some difference in performance of each edge-type, they all do significantly better than the null model.

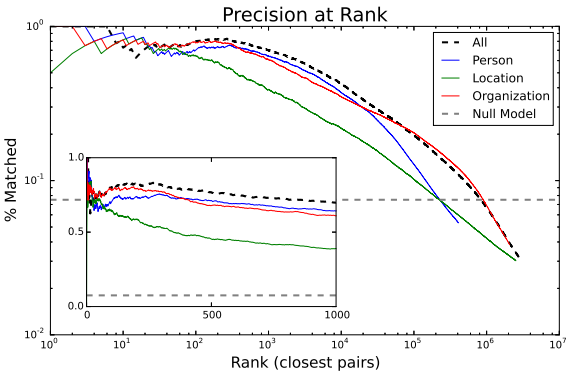

Fig. 6. The precision at rank for pairs of academics in decreasing order of similarity. Precision is defined as the pair being in the same department. In-set is a linear plot for the top 1,000 closest pairs.

To assess how useful the extracted NEs are as features in a statistical learning setting, they were used to train a multi-label classifier over departments. Table 1 shows the performance of the three feature-sets on a linear-kernel SVM and a random forest with 10 estimators [7, 9]. To make the comparison fair, the other models were restricted to using the same number of features (the number of edges in the graph, $|E \in G|$). Each model was trained in one-vs-all scheme according to each departmental label held by a faculty member (max = 5). Validation consisted of 20 folds on 20% held-out data. In all configurations, the node-incidence vectors, $\{E \in G\}$, outperformed the unigram models. Due to the multi-label scheme, the test-performance was averaged over all labels (for a single fold), which could falsely inflate the performance on held-out data. However, held-out performance was not our goal: the features in G were not mined for the purpose of training a classifier, nor was the weighting-scheme designed for such a task. Instead, we note its performance against relatively simple unigram models to underscore that G does indeed encode departmental structure, indicating it reliably represents academic relationships.

Feature-set	**Model**	**Precision**	**Recall**	**F_1Score**
$\{E \in G\}$	Linear-SVM	**0.53**	**0.95**	**0.66**
Count Unigrams	Linear-SVM	0.15	0.27	0.19
TF*IDF Unigrams	Linear-SVM	0.45	0.66	0.51
$\{E \in G\}$	Random Forest	**0.62**	**0.92**	**0.74**
Count Unigrams	Random Forest	0.60	0.89	0.69
TF*IDF Unigrams	Random Forest	0.61	0.92	0.71

Table 1. Multi-label classifier performance on three feature-sets: the weighted edge incidence vectors in G, count- and TF*IDF-weighted unigrams. Each set was used to train a linear-kernel support vector machine and a 10-estimator random forest. Reported are the test-scores averaged over 20 folds with randomized 80/20 train-test splits.

References

1. L. A. Adamic and E. Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.
2. M. Al Hasan, V. Chaoji, S. Salem, and M. Zaki. Link prediction using supervised learning. In *Workshop on Link Discovery: Issues, Approaches and Applications*, 2005.
3. A. Baronchelli, R. Ferrer-i Cancho, R. Pastor-Satorras, N. Chater, and M. H. Christiansen. Networks in cognitive science. *Trends in cognitive sciences*, 17(7):348–360, 2013.
4. S. Boccaletti, G. Bianconi, R. Criado, C. Del Genio, J. Gómez-Gardeñes, M. Romance, I. Sendiña-Nadal, Z. Wang, and M. Zanin. The structure and dynamics of multilayer networks. *Physics Reports*, 2014.

5. B. Bollobás. *Modern graph theory*, volume 184. Springer, 1998.
6. R. L. Breiger. The duality of persons and groups. *Social forces*, 53(2):181–190, 1974.
7. L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.
8. F. R. Chung. *Spectral graph theory*, volume 92. American Mathematical Soc., 1997.
9. C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
10. D. Duan, Y. Li, R. Li, and Z. Lu. Incremental k-clique clustering in dynamic social networks. *Artificial Intelligence Review*, 38(2):129–147, 2012.
11. R. I. Dunbar and M. Spoors. Social networks, support cliques, and kinship. *Human Nature*, 6(3):273–290, 1995.
12. D. Easley and J. Kleinberg. *Networks, crowds, and markets: Reasoning about a highly connected world*. Cambridge University Press, 2010.
13. S. L. Feld. The focused organization of social ties. *American journal of sociology*, pages 1015–1035, 1981.
14. J. R. Finkel, T. Grenager, and C. Manning. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 363–370. Association for Computational Linguistics, 2005.
15. J. Fischman. Arizona's big bet: The research rethink. *Nature*, 514(7522):292, 2014.
16. A. Gerow. Extracting clusters of specialist terms from unstructured text. In *Proceedings of the 2014 conference on Conference on Empirical Methods in Natural Language Processing (EMNLP 2014), Doha, Qatar*, pages 1426–1434, 2014.
17. A. Gerow and J. Evans. The modular community structure of linguistic predication networks. In *Proceedings of TextGraphs-9, Doha, Qatar*, pages 48–54, 2014.
18. L. Getoor and C. P. Diehl. Link mining: a survey. *ACM SIGKDD Explorations Newsletter*, 7(2):3–12, 2005.
19. L. Getoor, N. Friedman, D. Koller, and B. Taskar. Learning probabilistic models of link structure. *The Journal of Machine Learning Research*, 3:679–707, 2003.
20. K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. Eigentaste: A constant time collaborative filtering algorithm. *Information Retrieval*, 4(2):133–151, 2001.
21. R. Guns and R. Rousseau. Recommending research collaborations using link prediction and random forest classifiers. *Scientometrics*, 101(2):1461–1473, 2014.
22. B. Heintz and A. Chandra. Beyond graphs: toward scalable hypergraph analysis systems. *ACM SIGMETRICS Performance Evaluation Review*, 41(4):94–97, 2014.
23. P. W. Holland and S. Leinhardt. An exponential family of probability distributions for directed graphs. *Journal of the american Statistical association*, 76(373):33–50, 1981.
24. J. Lang and M. Lapata. Similarity-driven semantic role induction via graph partitioning. *Computational Linguistics*, 40(3):633–669, 2014.
25. D. Li, Z. Xu, S. Li, and X. Sun. Link prediction in social networks based on hypergraph. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 41–42, 2013.
26. L. Li and T. Li. News recommendation via hypergraph learning: encapsulation of user behavior and news content. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 305–314. ACM, 2013.
27. X. Li, B. Liu, and S. Y. Philip. Discovering overlapping communities of named entities. In *Knowledge Discovery in Databases: PKDD 2006*, pages 593–600. Springer, 2006.

28. D. Liben-Nowell and J. Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.
29. R. N. Lichtenwalter, J. T. Lussier, and N. V. Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.
30. H. Ma, H. Yang, M. R. Lyu, and I. King. Sorec: social recommendation using probabilistic matrix factorization. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 931–940. ACM, 2008.
31. P. Mika. Ontologies are us: A unified model of social networks and semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, 5(1):5–15, 2007.
32. R. Mitchum, A. Brand, and C. Transande. White paper: Information, interaction, influence: Research information technologies and their role in advancing science. 2014.
33. M. E. Newman. The structure and function of complex networks. *SIAM review*, 45(2):167–256, 2003.
34. A. Y. Ng, M. I. Jordan, Y. Weiss, et al. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 2:849–856, 2002.
35. G. Robins, P. Pattison, Y. Kalish, and D. Lusher. An introduction to exponential random graph (p^*) models for social networks. *Social networks*, 29(2):173–191, 2007.
36. A. Sharma, J. Srivastava, and A. Chandra. Predicting multi-actor collaborations using hypergraphs. *arXiv preprint arXiv:1401.6404*, 2014.
37. F. Shi, J. G. Foster, and J. Evans. Weaving the fabric of science: Dynamic network models of sciences unfolding structure. *Social Networks*, forthcoming, 2015.
38. S. Sintos and P. Tsaparas. Using strong triadic closure to characterize ties in social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1466–1475. ACM, 2014.
39. T. A. Snijders. Markov chain monte carlo estimation of exponential random graph models. *Journal of Social Structure*, 3(2):1–40, 2002.
40. C. Taramasco, J.-P. Cointet, and C. Roth. Academic team formation as evolving hypergraphs. *Scientometrics*, 85(3):721–740, 2010.
41. E. F. Tjong Kim Sang and F. De Meulder. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 142–147. Association for Computational Linguistics, 2003.
42. F. E. Walter, S. Battiston, and F. Schweitzer. A model of a trust-based recommendation system on a social network. *Autonomous Agents and Multi-Agent Systems*, 16(1):57–74, 2008.
43. F. Xia, Z. Chen, W. Wang, J. Li, and L. T. Yang. Mvcwalker: Random walk based most valuable collaborators recommendation exploiting academic factors. *IEEE Transcactions on Emerging Topics in Computing*, 2(3):364–375, 20.
44. Z.-K. Zhang and C. Liu. A hypergraph model of social tagging networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2010(10):P10005, 2010.