Link Partitioning on Simplicial Complexes
Using Higher-Order Laplacians

Xinyi Wu  
Institute for Data, Systems, and Society  
Massachusetts Institute of Technology  
Cambridge, USA  
xinyiwu@mit.edu

Arnab Sarker  
Institute for Data, Systems, and Society  
Massachusetts Institute of Technology  
Cambridge, USA  
arfabs@mit.edu

Ali Jadbabaie  
Institute for Data, Systems, and Society  
Massachusetts Institute of Technology  
Cambridge, USA  
jadbabai@mit.edu

Abstract—Link partitioning is a popular approach for discovering overlapping communities by identifying clusters of strongly connected links. Current link partitioning methods are specifically designed for networks modelled by graphs representing pairwise relationships. Therefore, these methods omit any higher-order information about group interactions in network data which is increasingly available. Simplicial complexes extend the dyadic model of graphs and can model polyadic relationships which are ubiquitous and crucial in many complex social and technological systems. In this paper, we introduce a link partitioning method that leverages higher-order (i.e. triadic and higher) information in simplicial complexes for better community detection. Our method utilizes a novel random walk on links of simplicial complexes defined by the higher-order Laplacian—a generalization of the graph Laplacian that incorporates polyadic relationships of the network. We transform this random walk into a graph-based random walk on a lifted line graph—a dual graph in which links are nodes while nodes and higher-order connections are links—and optimize for the standard notion of modularity. We show that our method is guaranteed to provide interpretable link partitioning results under mild assumptions. We also offer new theoretical results on the spectral properties of simplicial complexes by studying the spectrum of the link random walk. Experiment results on real-world community detection tasks show that our higher-order approach significantly outperforms existing graph-based link partitioning methods.

I. INTRODUCTION

Community detection is a central research topic in network science and has been mostly defined as a problem of determining groups of nodes who share dense connections [1]. However, because nodes often belong to multiple communities in networks, the task of community assignment is further complicated in practice [2]–[4]. Many overlapping community detection methods have been proposed to accommodate this problem [2]–[5]. In this paper, we focus on the “link community” paradigm, where communities are redefined as sets of closely interrelated links (edges) [2], [3], [5]. Unlike nodes, links in networks usually form for one dominant reason. This new definition of a community allows us to naturally find overlapping structures by redefining a node’s set of communities as incident link communities. Compared with node-based approaches, link partitioning has been found to reveal community structures with better quality in networks from various domains [2].

Current link partitioning algorithms are exclusively designed for graph representations of networks [2], [3], [5], which only model pairwise relationships between nodes. However, higher-order interactions in groups of more than two entities, which cannot be modelled by a graph, are ubiquitous and essential to understanding the structures and the behavior of complex systems in many fields [6]–[8] (Fig. 1). The complexity of real-world networks necessitates a higher-order informed link partitioning method that fundamentally captures polyadic relationships among nodes.

In this paper, we propose a link partitioning method that accounts for higher-order group behavior in networks. Our method models networks with higher-order (i.e. beyond dyadic) information using simplicial complexes and leverages tools from algebraic topology and combinatorial Hodge theory [9], [10]. In particular, our method derives link communities from link random walks based on higher-order Hodge Laplacians. This can be seen as a higher-order analogue of the use of spectral graph theory for node community detection, where one can interpret the action of normalized graph Laplacians as random walks on graphs. Motivated by the importance of triangular structures in social networks [7], [8] and their success in finding community structures in practice [11], we focus on triadic connections encoded by filled triangles, i.e. simplices of dimension 2, in simplicial complexes.

Inspired by a novel higher-order random walk recently introduced in Schaub et al. [12], we utilize the notion of a link-based random walk on oriented simplicial complexes to propose our new method for link partitioning. The random walk consists of what we denote as an upper walk, a lower walk, and a set of self-loops. The proposed link-based random walk can also be transformed into a weighted graph random walk on a lifted line graph \( \hat{G} \) where all oriented links become nodes, while nodes and filled triangles become links (Fig. 2). We take advantage of this transformation and use the transformed walk in conjunction with the Louvain algorithm to achieve our goal. Our method uses simplicial complexes to model both types of relationships and provides substantially better overlapping community detection results.
method [13] to design a modularity optimization algorithm for link partitioning. Since two orientations of each link are present in the lifted line graph $\hat{G}$, we show that under some mild conditions that generally hold for real networks, the proposed algorithm will always cluster the two orientations of each link into the same community. That is, the community partition in the lifted line graph $\hat{G}$ has a direct interpretation as link communities in the original higher-order network.

In addition to the algorithm itself, we connect to the spectral theory for simplicial complexes by proving a set of fundamental spectral properties for the lifted line graph random walk $\hat{P}$. We identify the equivalence between the link random walk and a diffusion process on the simplicial complex propagating via the higher-order Laplacian [14], [15]. Our results show that the spectrum of the random walk matrix $\hat{P}$ consists of two parts, which we call the “even” part and the “odd” part. The even part has symmetry in the space on the lifted line graph $\hat{G}$, whereas the odd part corresponds to the spectrum of the Laplacian and thus it is connected to the homology of the simplicial complex. There have been constant developments over years building a spectral theory for simplicial complexes [14]–[16]. Our work provides a new venue for understanding the spectral properties of simplicial complexes through the lifted line graph $\hat{G}$ and the random walk $\hat{P}$.

Finally, we test our proposed method in seven real-world networks where higher-order information is available. On average, our method has 12% improvement over the graph baselines. The comparisons suggest that higher-order information is valuable and thus our higher-order framework overall provides better solutions.

In summary, our paper develops a link partitioning method that incorporates higher-order information in networks. Theoretically, the proposed method generalizes the use of graph Laplacian for community detection to higher-order network data structure. It also provides new insights into the spectral theory for simplicial complexes. Algorithmically, the method takes advantage of the lifted line graph transformation and the Louvain method for community detection. We show that optimizing modularity in the lifted line graph using the Louvain method is guaranteed to produce interpretable link partitioning results under realistic conditions. Empirically, we find that our higher-order method results in substantial improvements in discovering overlapping community structure. Our work uses simplicial complexes to exploit rich and valuable higher-order network data and opens up new possibilities for higher-order network problems that apply across a broad set of domains.

II. PRELIMINARIES

A. Modularity from a Random Walk Perspective

To motivate our method of using a random walk on links for link partitioning, let us first consider the well-known concept of modularity [17] in terms of random walks on nodes. Given an undirected (weighted) graph $G$ with adjacency matrix $A$, the modularity $Q$ used for evaluating quality of a node partition $P$, is defined as $Q = \frac{1}{2m} \sum_{C \in P} \sum_{i,j \in C} [A_{ij} - \frac{k_{i,j}}{2m}]$, where $k_{i} = \sum_{j} A_{ij}$ is the sum of the weights of the links attached to node $i$, $2m = \sum_{i,j} A_{i,j}$ is the sum of all of the link weights in the graph, and $C'$ runs over all the communities in $P$.

A random walk interpretation of the modularity $Q$ is the follows [3], [18]: denote the probability of a random walker on node $i$ at time step $t$ as $p_{i,t}$, where the dynamics are given by the standard unbiased random walk $p_{i,t+1} = \sum_{j} \frac{A_{ij}}{\sum_{l} A_{il}} p_{i,t}$. When the given network is undirected, connected and non-bipartite, one can show that the stationary solution of the dynamics is $p_{i} = k_{i}/2m$. Consider a community $C \in P$. If the system is at equilibrium, the probability that a random walker stays in $C$ during two successive time steps is $\sum_{m,n \in C} \frac{k_{m,n}}{2m}$, whereas the probability of finding two independent walkers in $C$ is $\sum_{i,j \in C} \frac{k_{i,j}}{2m^{2}}$. Hence one can reinterpret $Q$ as a summation over the communities of the difference of the two probabilities. This interpretation suggests a natural generalization of modularity that allows one to tune its resolution as follows: one can generalize $Q$ to paths of arbitrary length $t$ as $Q(A,t) = \frac{1}{2m} \sum_{C \in P} \sum_{i,j \in C} (T^{t})_{ij} k_{ij} - \frac{k_{i,j}}{2m}$, where $T_{ij} = A^{t}_{ij} k_{j}$ [18].

The above random walk formulation of modularity suggests that one should look at a random walk process traversing the links of a network to find communities of links. In order to define a link random walk on a higher-order network, we introduce simplicial complexes and higher-order Hodge Laplacians for higher-order network modelling.

B. Simplicial Complexes

Let $V$ be a finite set of nodes. A $k$-simplex $S^{k}$, or a simplex of dimension $k$, is a subset of $V$ with $k + 1$ elements. A simplicial complex (SC) $X$ is a set of simplices with the property that if $S \in X$, then all subsets of $S$ are also in $X$. We use $X^{k}$ to denote the subset of all $k$-simplices in $X$. A $k$-simplex has $k + 1$ subsets of dimension $k - 1$, which are called faces. If $S^{k-1}$ is a face of simplex $S^{k}$, then $S^{k}$ is called a co-face of $S^{k-1}$ [10].

Two $k$-simplices in an SC $X$ are upper adjacent if they are both faces of the same $(k + 1)$-simplex and are lower adjacent if both share a common face. For any simplex in $X$, we define its degree, denoted by $deg(x)$, to be the number of its co-faces.

We endow arbitrarily each simplex with an orientation. An orientation can be seen as a chosen ordering of the elements of a simplex, modulo even permutations. For simplicity, we choose the reference orientation of the simplices induced by the node labels $\{s_{1}^{k} = [i_{0}, ..., i_{k}], i_{0} < ... < i_{k}\}$.

C. Boundary Operators and Hodge Laplacians

For each dimension $k$, we define the finite-dimensional vector space $C_{k}$ with coefficients in $\mathbb{R}$, whose basis elements are the oriented simplices $s_{k}^{i}$. An element $c_{k} \in C_{k}$ is called a $k$-chain, and is a linear combination of the basis elements $c_{k} = \sum a_{i} s_{k}^{i}$. Thus, each element $c_{k} \in C_{k}$ can be represented by a vector $a = [a_{1}, ..., a_{n_{k}}]$, where $n_{k} = |X^{k}|$ is the number of $k$-simplices in $X$. A change of the orientation of the basis $s_{k}^{i}$ is defined to be a flip of the sign of its coefficient $a_{i}$.

Given the space of chains $C_{k}$, we define the linear boundary maps $\partial_{k} : C_{k} \rightarrow C_{k-1}$ [10] as $\partial_{k}(\{i_{0}, ..., i_{k}\}) = ...$
generalization of the relationship between the normalized graph Laplacian and the random walk on graph [19]. To state our results compactly, we define the following matrices: 
\[
B_1 := B_1 V^T = [B_1 - B_1] , \quad B_2 := VB_2 = [B_2 - B_2].
\]
We denote \(B_1^+\) and \(B_1^-\) as the positive and the negative part of these matrices, respectively, i.e. \((B_1^+)_ab = \max\{(B_1)_ab,0\}, (B_1^-)_ab = \max\{(-B_1)_ab,0\}.

**Theorem 1.** The matrix \((I - L_1)/2\) has a stochastic lifting, i.e. there exists a stochastic matrix \(\tilde{P}\) such that \((I - L_1)V^T = 2V^T \tilde{P}\), where \(\tilde{P}\) corresponds to a random walk on an undirected lifted line graph \(\hat{G}\) with adjacency matrix
\[
\hat{A} = \hat{A}^1 + \hat{A}^a + \hat{A}^n , \quad \hat{A}^1 = [B_1^+]^T [B_1^+] + [B_1^-]^T [B_1^-], \quad \hat{A}^a = [B_2^+]^T [B_2^-] + [B_2^-]^T [B_2^-] , \quad \text{and} \hat{A}^n is a diagonal matrix with \(|\hat{A}^n|_{ij,ij} = \deg(i) + \deg(j) + 3\deg(i,j).\)

The proof is in Appendix A. In the lifted line graph \(\hat{G}\), nodes are oriented links of \(\mathcal{X}\), while links are nodes and filled triangles of \(\mathcal{X}\). Figure 2 gives an illustration of the lifting process, the lifted line graph \(\hat{G}\), and the three components of \(\hat{A}\). One can interpret \(\hat{A}^1\) as describing the connections between lower adjacent links that are aligned with respect to the reference orientations, while \(\hat{A}^n\) as describing the connections between upper adjacent links that are not aligned respect to the reference orientations. Finally, by construction, the random walk \(\tilde{P}\) contains self-loops encoded by \(\hat{A}^a\).

If we denote \(d_{\hat{G}}(i,j) = 2(\deg(i) + \deg(j) + 3\deg(i,j))\), \(\hat{P}\) can be interpreted as follows: starting from state \([i,j]\), with probability 1/2, the walker remains in \([i,j]\); with probability \((\deg(i) + \deg(j))/d_{\hat{G}}(i,j)\), the walker transitions to a lower adjacent state (lower walk); with probability \(3\deg(i,j)/d_{\hat{G}}(i,j)\), the walker transitions to an upper adjacent state (upper walk).

Moreover, the adjacency matrix \(\hat{A}\) has a special block form:

**Proposition 1.** The adjacency matrix \(\hat{A}\) of \(\hat{G}\) is symmetric and has the following block form:
\[
\hat{A} = \begin{bmatrix} \hat{A}_I & \hat{A}_{IJ} \\ \hat{A}_{JI} & \hat{A}_J \end{bmatrix}.
\]

**B. Modularity Optimization Using the Louvain Method**

Having defined \(\tilde{P}\) and \(\hat{A}\), we plug \(\hat{A}\) into the definition of modularity (Section II-A) to find the optimal link partitions in terms of modularity in \(\hat{G}\) with varying resolution. In practice, such optimal partitions can be found by standard modularity optimization algorithms. One of the most popular and state-of-the-art modularity optimization methods is the Louvain method [13], which has two phases. In the first phase, one starts by putting each node in the graph into a distinct community. Then the algorithm moves each node \(i\) to the community of some neighbor \(j\) that yields the largest gain in \(\Delta Q\). This process repeats until moving nodes can no longer improve the modularity. Then we enter the second phase, where we collapse the existing communities to single nodes and build a new network where nodes are the communities from the previous phase, and links weights are aggregated accordingly. Once the new network is created, the entire process is repeated on the new network.

Since we include two orientations of each link as nodes in the lifted line graph \(\hat{G}\), there is a possibility that applying a
modularity optimization method directly to $\tilde{G}$ could break the two orientations of a same link into different groups. Such a partitioning result would be hard to interpret. However, the following theorem shows this is generally not the case.

**Theorem 2.** Given the lifted line graph $\tilde{G}$ with adjacency matrix $A$, let $2\tilde{m} = \sum_{i,j} A_{ij}$. Let $k_i = \sum_j A_{ij}$ be the sum of the weights of the connections attached to the oriented link $i$. Suppose that for every node in $\tilde{G}$, $k_i \leq \sqrt{2\tilde{m}}$ (*). In addition, assume in the first phrase, we iterate through links with the chosen orientations $\{e_0, e_1, ..., e_N\}$ first, and then iterate through the links with the reverse orientations $\{\bar{e}_0, \bar{e}_1, ..., \bar{e}_N\}$ . Then the Louvain method $A$ will always cluster two orientations of a link $e$ and $\bar{e}$ into the same community, i.e. $A(e) = A(\bar{e})$.

The proof is in Appendix A. Here (*) is a sufficient but not necessary condition for the theorem. We verified that the networks in our experiments (Section IV) indeed satisfy (*) and thus Theorem 2 applies. One can deliberately construct counterexamples that violate (*), but this condition generally holds for large realistic networks.

**Corollary 1.** Let $S(\tilde{G})$ be the line graph by considering the two orientations of the same link in $\tilde{G}$ as one supernode, aggregating the link weights accordingly. Then $S(\tilde{G})$ has adjacency matrix $A_1 + A_{11}$. Under (*), optimizing modularity on $S(\tilde{G})$ with the Louvain method is equivalent to optimizing modularity on $\tilde{G}$ with the Louvain method.

**C. Spectral Properties of the Stochastic Lifting Matrix $\hat{P}$**

Our assumption about the graph skeleton of $X$ being undirected, connected and non-bipartite implies that $\tilde{G}$ is also undirected, connected and non-bipartite. Hence $P$ is reversible [20]. The reversibility implies that $P$ is diagonalizable and has a real spectrum. We state the following spectral properties about $P$.

**Proposition 2.** $P$ has the following block form: $P = \begin{bmatrix} \hat{P}_1 & \hat{P}_{11} \\ \hat{P}_{11} & \hat{P}_1 \end{bmatrix}$.

**Proposition 3.** Let $\text{Spec}(M)$ denote the spectrum of a matrix $M$. Then $\text{Spec}(P) = \text{Spec}(\hat{P}_1 + \hat{P}_{11}) \cup \text{Spec}(\hat{P}_1 - \hat{P}_{11})$.

**Proposition 4.** If $x$ is an eigenvector for $\hat{P}_1 + \hat{P}_{11}$, then $[x, x]^T$ is an eigenvector for $P$ with the same eigenvalue.

**Corollary 2.** 1 is an eigenvalue for $\hat{P}_1 + \hat{P}_{11}$, and suppose its corresponding eigenvector is $x$. Then 1 is an eigenvalue for $P$ with eigenvector $[x, x]^T$.

**Corollary 3.** Let the stationary solution of $\hat{P}$ be $\hat{\pi}$. Then $V^T \hat{\pi} = 0$.

The above properties of $\hat{P}$ imply that one can think of the random walk on the lifted line graph $\tilde{G}$ as composed of two parts: the even part $P_1 + P_{11}$ and the odd part $P_1 - P_{11}$. We call the sum “even” because it suggests the symmetry between the two orientations in $\tilde{G}$. As shown in Proposition 4, the eigenvalues of $\hat{P}$ that come from the even part have eigenvectors that put the same value on both orientations for each link. Meanwhile, the difference is named “odd” because if one wishes to study a higher-order diffusion dynamic on SCs, one should consider the “expectation process” defined in [14], or equivalently, the “lift-propagate-project” procedure given by the definition of lifting: given link flows defined on $X$, we first lift them to $\tilde{G}$ via $V$, then the state is propagated through $\hat{P}$, and finally we project the lifted flows to the space on $X$ via $\frac{1}{2}V^T$ [12]. The difference is exactly what remains after the projection, as the even part will cancel each other during the projection.

**D. Method Summary and Computational Complexity**

Algorithm 1 summarizes the proposed method. Given a SC $X$, computing $B_1 \in \mathbb{R}^{n_0 \times n_1}$ and $B_2 \in \mathbb{R}^{n_1 \times n_2}$ requires runtime $O(n_1)$ and $O(n_2)$, respectively and building $\hat{A}$ according to the formulas in Theorem 1 runs in time $O(n_1 + n_2)$. Moreover, the Louvain method is observed in practice to run in time $O(n_1 \log n_1)$ on average [13]. So the overall computational complexity of Algorithm 1 is $O(n_1 \log n_1 + n_2)$. Compared with running the Louvain method on a graph, the extra term $n_2$ is a trade-off between higher-order information and computational efficiency.

**Algorithm 1.** Higher-order link partition using $L_1$ and lifting

**Input:** a SC network $X$, the Louvain method $A$.

1. Compute boundary maps $B_1$ and $B_2$ from $X^1$ and $X^2$.
2. Build the adjacency matrix $\hat{A}$ for the lifted line graph $\tilde{G}$.
3. Feed $\hat{A}$ to $A$ and obtain the community structure $C$ of $\tilde{G}$.
4. Under (*), for $e \in X^1$, $A(e) = A(\bar{e})$. $C$ can be directly projected to $X$ and be interpreted as communities of links in $X$.

**Output:** Link communities of $X$ (communities of $X^1$).

**IV. EXPERIMENTS**

In this section, we evaluate the performance of our algorithm on real-world network data with higher-order information. In all the experiments, we compare our method (denoted by $\hat{A}$) with two well-known types of graph-dimensional methods for link-based community detection:

1. Code is available at https://github.com/xinyiwu98/SC_link_comm.
Dendrogram cutting $S$: Ahn et al. propose a dendrogram cutting method, which we denote as $S$. It is built on a similarity score and an objective called partition density [2].

Modularity based approaches $C$, $D$, and $E_{i}$: Evans and Lambiotte propose modularity optimization on three different weighted line graphs (corresponding adjacency matrices $C$, $D$ and $E_{i}$) [3], and we use the Louvain method for the actual optimization in the experiments.

A. Datasets and Evaluation Measures

We focus on social networks where filled triangles have expressive power and clear interpretations [7], [8]. Summary statistics of the datasets used are present in Table I. See Appendix B for detailed descriptions for each dataset and the metadata used accordingly. For each dataset, we examined the same 100 networks whose sizes ranged between 59 – 93 nodes and reported the average results.

| Datasets Used in the Experiments |
|----------------------------------|
| Name                            | Nodes | Links | Triangles |
| contact-high-school [21]         | 327   | 5,818 | 2,370     |
| contact-primary-school [22]      | 242   | 8,317 | 5,139     |
| email-Enron [23]                 | 144   | 1,344 | 1,359     |
| email-Eu [23]                    | 979   | 29,299| 160,605   |
| congress-bills [23], [24]        | 1712  | 66,102| 86,164    |
| senate-committees [25], [26]     | 282   | 14,224| 169,362   |
| house-committees [25], [26]      | 1290  | 126,155| 282       |

As Ahn et al. point out, there are multiple aspects that one must consider when comparing community detection algorithms. Hence, we consider four distinct metrics proposed in [2] to benchmark our approach. For each metric, the higher the value, the better the partitioning result on what it measures.

Community quality Each dataset possesses metadata that describe each node. We assume that nodes in the same communities share more metadata than nodes in different ones. We then compute the enrichment score of node pair: $\text{Enrichment} = \frac{s(i,j)|\text{metadata community}}{s(i,j)|\text{par. group}}$, where $s(i,j)$ is a metadata-based similarity score between node $i$ and $j$, and $\gamma$ denotes the average over the particular group.

Overlap quality Overlap quality measures how much information about the true overlap of nodes is learned from the link communities. We calculate the mutual information between the number of memberships of each node given by the link communities and the metadata closely related to the number of true communities that each node participates in.

Community coverage Community coverage measures how much of the network is studied. We count the fraction of nodes that belong to at least one nontrivial community given by the link partition, i.e. community with $\geq 3$ nodes.

V. DISCUSSION

In this paper, we use a novel random walk on links of simplicial complexes to perform link partitioning that incorporates higher-order network information. We find that using higher-order information provides substantial improvements in discovering overlapping community structure through links. These results emphasize the value of higher-order network data for studying the structure and behavior of networks and necessitate higher-order network models that can take account of this information. Since our method connects to the spectral
theory for simplicial complexes through random walks, future research should aim to study this relationship more.

VI. ACKNOWLEDGEMENT

Xinyi Wu would like to thank Zihui Wu for inspiring discussions. This research has been supported in part by ARO grant number W911NF-19-1-0057, and a Vannevar Bush Fellowship from the Office of the Secretary of Defense.

REFERENCES

[1] S. Fortunato, “Community detection in graphs,” Physics Reports, 2010.
[2] Y.-Y. Ahn, J. P. Bagrow, and S. Lehmann, “Link communities reveal multiscale complexity in networks,” Nature, 2010.
[3] T. S. Evans and R. Lambiotte, “Line graphs, link partitions, and overlapping communities,” Phys. Rev. E Stat. Nonlin. Soft Matter Phys., 2009.
[4] G. Palla, I. Derényi, I. J. Farkas, and T. Vicsek, “Uncovering the overlapping community structure of complex networks in nature and society,” Nature, 2005.
[5] T. S. Evans and R. Lambiotte, “Line graphs of weighted networks for overlapping communities,” The European Physical Journal B, 2010.
[6] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. B. Chklovskii, and U. Alon, “Network motifs: simple building blocks of complex networks.” Science, 2002.
[7] P. Holland and S. Leinhardt, “A method for detecting structure in sociometric data,” Am. J. Sociol., 1970.
[8] M. S. Granovetter, “The strength of weak ties,” Am. J. Sociol., 1973.
[9] L.-H. Lim, “Hodge laplacians on graphs,” ArXiv, vol. 1507.05379, 2020.
[10] A. Hatcher, “Algebraic topology,” 2002.
[11] A. R. Benson, D. F. Gleich, and J. Leskovec, “Higher-order organization of complex networks.” Science, 2016.
[12] M. T. Schaub, A. R. Benson, P. Horn, G. Lipper, and A. Jadbabaie, “Random walks on simplicial complexes and the normalized hodge 1-laplacian,” SIAM Review, 2020.
[13] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, “Fast unfolding of communities in large networks,” J. Stat. Mech., 2008.
[14] O. Parzanchevski and R. Rosenthal, “Simplicial complexes: Spectrum, homology and random walks,” Random Struct. Algorithms, 2017.
[15] S. Mukherjee and J. Steenbergen, “Random walks on simplicial complexes and harmonics,” Random Struct. Algorithms, 2016.
[16] T. Kaufman and I. Oppenheim, “Higher order random walks: Beyond spectral gap,” Combinatorica, 2018.
[17] M. E. J. Newman, “Modularity and community structure in networks.” PNAS, 2006.
[18] J.-C. Delvenne, S. N. Yaliraki, and M. Barahona, “Stability of graph communities across time scales,” PNAS, 2010.
[19] F. R. K. Chung, “Spectral graph theory,” 1996.
[20] D. Aldous and J. A. Fill, “Reversible markov chains and random walks on graphs.” 2002.
[21] R. Mastrandrea, J. Fournet, and A. Barrat, “Contact patterns in a high school: A comparison between data collected using wearable sensors, contact diaries and friendship surveys,” PLoS ONE, 2015.
[22] J. Stehlé, N. Voirin, A. Barrat, C. Cattuto, L. Isella, J.-F. Pinton, M. Quaggiotti, W. V. den Broeck, C. Régis, B. Lina, and P. Vanhems, “High-resolution measurements of face-to-face contact patterns in a primary school.” PLoS ONE, 2011.
[23] A. R. Benson, R. Abebe, M. Schaub, A. Jadbabaie, and J. Kleinberg, “Simplicial closure and higher-order link prediction.” PNAS, 2018.
[24] J. H. Fowler, “Connecting the congress: A study of cosponsorship networks,” Political Analysis, 2006.
[25] P. S. Chodrow, N. Veldh, and A. R. Benson, “Generative hypergraph clustering: From blockmodels to modularity.” Science Advances, 2021.
[26] C. Stewart and J. Woon, “Congressional committee assignments, 103rd to 105th congresses, 1993–1998, data,” 2005.
[27] L. Peel, D. B. Larremore, and A. Clauset, “The ground truth about metadata and community detection in networks,” Science Advances, 2017.

APPENDIX A

PROOFS

Proof of Theorem 1. First, \( -L_1 V^T = V^T (\tilde{A}^T + \tilde{A}^v) \) (12), Lemma 3.2). Then observe that by definition, \( \tilde{P} = \text{Adj}(\tilde{A})^{-1} = \frac{1}{2} \tilde{A} \left[ \begin{array}{cc} D_{tot} & 0 \\ 0 & D_{tot} \end{array} \right] \) and \( \tilde{A}^v = \left[ \begin{array}{cc} D_{tot} & 0 \\ 0 & D_{tot} \end{array} \right]. \) So \( D_{tot} V^T = V^T \tilde{A}^v. \) It follows that \( V^T \tilde{P} = (D_{tot} - L_1) V^T (\tilde{A})^{-1} = (D_{tot} - L_1) \frac{1}{2} (I - L_1) V^T. \)

Proof of Theorem 2. Consider the beginning state, where all the links are in their own communities. Then putting oriented link \( e_0 \) into the community \( \tilde{C} \) will incur a change in modularity \( Q \) by \( \Delta Q(e_0 \rightarrow \tilde{C}) = \sum_{e \notin C} \sum_{e \in \tilde{C}} k_e \min(A_{e_0 e}, A_{e \tilde{C}}) \geq 0 \) under \( \ast \). For any other neighboring nodes \( e' \) for \( e_0 \), consider \( \Delta Q(e_0 \rightarrow \tilde{C}) = \sum_{e \notin C} \sum_{e' \in \tilde{C}} k_e \min(A_{e_0 e}, A_{e e'}) \geq 0 \). Notice that \( K_{e_0 e} \) maximizes with \( e = e' = e_0 \), such that \( K_{e_0 e} = 2 (\deg(e_0)) \) and the difference with other \( K_{e_0 e'} \) is at least 2, while \( (k_{e_0} - k_{e_0}) k_{e_0} \leq 2m^2 \). Hence \( \Delta Q(e_0 \rightarrow \tilde{C}) = \sum_{e \notin C} \sum_{e' \in \tilde{C}} \min(A_{e_0 e}, A_{e e'}) \\ \geq 0, \mbox{\mbox{\mbox{veil}}} \). The Louvain method will put \( e_0 \) into the community of \( e_0 \).

Then consider \( e_1 \). Let the community of \( e_0 \) and \( \tilde{C} \) be \( \tilde{C}_0 \). Based on a similar reasoning as above, it suffices to show that the Louvain method will put \( e_1 \) in the community of \( e_0 \) not \( \tilde{C}_0 \). Notice that \( \Delta Q(e_1 \rightarrow \tilde{C}_1) = \sum_{e \notin C} \sum_{e' \in \tilde{C}} k_e \min(A_{e_0 e}, A_{e e'}) \geq 0 \). We have \( K_{e_1, e_0} = 2 (\deg(e_1)) \), and \( K_{e_1, e_0} \leq 2 (1 + \deg(e_1)) \). So \( \Delta Q(e_1 \rightarrow \tilde{C}_1) = \sum_{e \notin C} \sum_{e' \in \tilde{C}} \min(A_{e_0 e}, A_{e e'}) \\ \geq 0, \mbox{\mbox{\mbox{veil}}} \). We use the department memberships (email-Eu) and job positions (email-Enron) as our node similarity measure, and the total number of emails that each email address sent as our proxy for overlap quality.

APPENDIX B

DATA DESCRIPTION

Here are detailed descriptions of the seven real-world datasets used in Section IV. The choices of the metadata carefully followed the choices of metadata in Ahn et al. [2].

- **contact-primary-school, contact-high-school:** nodes are individuals, and simplices form if individuals email one another. We used the classification of emails for each individual as our proxy for overlap quality.
- **email-Eu, email-Enron:** nodes are email addresses and a simplex forms if a set of individuals email one another. We used the department memberships (email-Eu) and job positions (email-Enron) as our node similarity measure, and the total number of emails that each email address sent as our proxy for overlap quality.
- **congress-bills, senate-committees, house-committees:** nodes are members of Congress and a simplex forms if a set of members co-sponsored a bill or are part of the same committee. We used political party affiliation as our similarity measure, and the number of elected terms as our overlap quality proxy.