OPTANE: An OPtimal Transport Algorithm for NEtwork Alignment

Farzan Masrour
Michigan State University
East Lansing, MI, USA
masrours@cse.msu.edu

Pang-Ning Tan
Michigan State University
East Lansing, MI, USA
ptan@cse.msu.edu

Abdol-Hossein Esfahanian
Michigan State University
East Lansing, MI, USA
esfahanian@cse.msu.edu

***Abstract*—Networks provide a powerful representation tool for modeling dyadic interactions among interconnected entities in a complex system. For many applications such as social network analysis, it is common for the entities to appear in more than one network. Network alignment (NA) is an important first step towards learning the entities' behavior across multiple networks by finding the correspondence between similar nodes in different networks. However, learning the proper alignment matrix in noisy networks is a challenge due to the difficulty in preserving both the neighborhood topology and feature consistency of the aligned nodes. In this paper, we present OPTANE, a robust unsupervised network alignment framework, inspired from an optimal transport theory perspective. The framework provides a principled way to combine node similarity with topology information to learn the alignment matrix. Experimental results conducted on both synthetic and real-world data attest to the effectiveness of the OPTANE framework compared to other baseline approaches.**

***Index Terms*—Network Alignment, Optimal Transport.**

I. Introduction

Networks are powerful tools for modeling interactions between entities in a complex system. Examples include the Internet as a physical network of interconnected computing devices, social media platforms such as Facebook for human communications, and genetic regulatory networks in biochemical systems. Analyzing properties of these networks may provide useful insights into the underlying mechanism governing the behavior of the complex system. With the proliferation of data collected in many of these application domains, it is becoming increasingly common for the entities to appear in more than one network. For instance, there are many individuals who are members of multiple social networks. Similarly, the same protein may participate in different interaction networks across species. Previous studies have demonstrated the utility of analysing the multiple networks simultaneously, from providing better recommendation on e-commerce websites [1] to improving protein function prediction [2]. However, as the first step towards enabling such analysis, one must be able to match or link together nodes from different networks. Network alignment seeks to aid this undertaking by learning the mapping between nodes in multiple networks based on their neighborhood topology and feature similarity information.

Finding the correspondence between nodes in two networks is akin to the subgraph isomorphism problem, which is known to be NP-Complete [3]. Numerous algorithms have been developed in recent years to address this computational challenge. Most of the algorithms are proposed to preserve the following requirements: (1) **Topological consistency**, i.e., the aligned nodes must have similar neighborhood structure across different networks. A popular network alignment method based on topological consistency is isoRank [4], which recursively defines the alignment between nodes in different networks based on similarity of their neighbors. (2) **Feature consistency**, which refers to the property that the aligned nodes must share similar node or edge features. FINAL [5] is an example of an algorithm that utilizes both the node and edge features to learn the alignment between nodes in two networks.

Despite the extensive research, current approaches are still ineffective when applied to networks in which the link structure and node features are noisy. To address this challenge, this paper introduces OPTANE, a novel framework for unsupervised network alignment, inspired by ideas from Optimal Transport (OT) theory for comparing data with different distributions [6]. OT has received considerable attention in recent years due to its successes in numerous fields, from economics to image processing, language processing, and generative models in machine learning [7]. In this paper, we design an OT approach to map the nodes from one network to another in such a way that minimizes a transportation cost function. Specifically, we leverage the differences in the node features to calculate the transportation cost between the aligned nodes in different networks.

One potential challenge in applying OT to the network alignment problem is computational efficiency as the complexity for solving a discrete OT problem using finite-dimensional linear program is $O(n^3log(n))$ [8]. A standard way to address this challenge is to compute a regularized version of OT using entropy regularization [9], which allows the use of more efficient methods such as the Sinkhorn algorithm to solve the optimization problem. Instead of using entropy regularization, OPTANE considers a *topological regularization* approach to learn the OT matrix by accounting for the difference in neighborhood structure between the aligned nodes. This regularization approach allows us to design a projected gradient descent approach to efficiently learn the mapping between nodes while preserving their topological consistency.

Empirical results showed that OPTANE outperformed several baseline methods when applied to networks with perturbed node and link structure.

II. Related Work

In principle, network alignment can be formulated as an integer quadratic programming problem [10], which unfortunately is a computationally hard problem to solve. Various approximation methods were proposed to overcome this problem. For example, Bayati et al.[10] employed a message passing algorithm based on belief propagation while Klau [11] converted it into a linear programming problem. Other binary relaxation approaches proposed include those by Koutra et al. [12], Zhang and Tong [5] and Vogelstein et al. [13]

IsoRank by Singh et al. [4] is one of the most well-known network alignment algorithm. IsoRank iteratively learns the alignment score between two nodes based on the scores of its neighbors. Variations of IsoRank include the works by Kollias et al. [14] and Liao et al. [15], to reduce computational cost and improve interpretability of similarity scores.

III. Preliminaries

A. Network Alignment

There are various formulations to the network alignment problem. Here we consider the global network alignment problem, which seeks to find a mapping between the nodes in multiple networks. In contrast, the local network alignment problem aims to find multiple similar sub-networks between two networks. The global network alignment problem can be formally stated as follows: Let $G^{(1)} = (V^{(1)}, E^{(1)}, F^{(1)})$ and $G^{(2)} = (V^{(2)}, E^{(2)}, F^{(2)})$ be a pair of attributed networks, where $V^{(i)}$ is the set of nodes in $G^{(i)}$, $E^{(i)}$ is its corresponding links, and $F^{(i)}$ is its node attribute matrix. Furthermore, let $n_1 = |V^{(1)}|$ and $n_2 = |V^{(2)}|$. The global network alignment task is to learn an $n_1 \times n_2$ matrix T, where T_{ij} denotes the alignment (coupling) score between nodes $v_i \in V^{(1)}$ and $u_j \in V^{(2)}$.

B. Optimal Transport

Optimal Transport (OT) provides a useful geometry for comparing probability distributions. The original OT problem was proposed by Monge [6] in 1781. Its modern formulation as a linear program is attributed to Kantorovich [16]. The discrete version of OT can be cast as a linear programming problem:

$$\begin{gathered} argmin_{T \in U(p,q)} f(T) = < C, T >_F \\ U(p,q) = \Big\{ T \in \mathcal{R}_+^{n_1 \times n_2} | T\mathit{1}_{n_2} = q, T^\top \mathit{1}_{n_1} = p \Big\} \end{gathered} \tag{1}$$

The above problem can be interpreted as finding an $n_1 \times n_2$ transport matrix T between two sets of points, $X = \{x_1, \ldots, x_{n_1}\}$ and $Y = \{y_1, \ldots, y_{n_2}\}$ such that the overall mapping cost with respect to an $n_1 \times n_2$ cost matrix C is minimized. Since it is a linear program with $O(n)$ linear constraints, it can be solved using standard linear programming method in $O(n^3 log(n))$ [17].

IV. Proposed Framework

This section describes our proposed framework using OT for network alignment. Given two attributed networks, $G^{(1)}$ and $G^{(2)}$, our goal is to learn a transport matrix T between $V^{(1)}$ and $V^{(2)}$ in a way that minimizes the cost matrix C. In this work, we define the cost matrix based on the distance between their node features $F^{(1)}$ and $F^{(2)}$:

$$c(x_i, y_j) = dist(F^{(1)}[i,:], F^{(2)}[j,:])$$

In the context of network alignment, the probability distributions p and q given in Eqn. (1) can be interpreted as a measure of relative importance of the nodes in the network. Here, we assume a uniform distribution over $V^{(1)}$ and $V^{(2)}$. Furthermore, standard OT formulation typically incorporates entropy regularization to induce specific properties to the solution. For network alignment, we consider the following topological and sparsity regularizers.

Topological regularizer, whose objective is to preserve the neighborhood structure of the transportation process. In other words, if a pair of nodes are linked to each other in the first network, it should also be linked in the second network after transportation. One way to achieve this is by introducing the following reqularization penalty:

$$\Omega_{to}(T) = ||A^{(1)}T - TA^{(2)}||_F^2$$

Sparsity regularizer, whose objective is to ensure that each node in the second graph is aligned to only one node in the first graph (assuming $n_2 \leq n_1$).

$$\Omega_{sp}(T) = ||T^\top T - I||_F^2$$

Putting it altogether, the objective function for the OPTANE framework is given as follows:

$$argmin_{T \in U(p,q)} < C, T >_F + \frac{\mu}{2}\Omega_{to}(T) + \frac{\alpha}{2}\Omega_{sp}(T) \tag{2}$$

The constraint $U(p,q)$ is a bounded, convex polytope defined by $n_1 + n_2$ equality constraints. For the network alignment problem, we replace $U(p,q)$ with

$$B = \Big\{ T \in \mathcal{R}^{n_1 \times n_2} | 0 \leq T_{ij} \leq 1 \Big\}$$

and employ a projected gradient descent algorithm to solve the constraint optimization problem. The projected gradient descent includes the following two steps:

1) Update T based on the following gradient term of the objective function:

$$\mu\Big({A^{(1)}}^2 T - 2A^{(1)}TA^{(2)} + T{A^{(2)}}^2\Big) + C + 2\alpha T\Big(T^\top T - I\Big)$$

2) Project T to B by mapping all negative entries to zero and all values larger that 1 to 1.

V. Experimental Evaluation

This section presents the experiments conducted to evaluate the performance of the proposed OPTANE framework.

	n_1	n_2	# node pairs
Douban(Online/offline)[5]	3,906	1,118	1,118
Flickr/Lastfm [5]	15,436	12,974	452
Flickr/myspace [5]	10,733	6,714	267
Deezer/Deezer [20]	12,286	10,490	5655
facebook [21]	4,039	3,992	3,992

Table I
SUMMARY STATISTICS OF REAL-WORLD NETWORK DATA.

A. Data

1) Synthetic Data: We employed the Barabási-Albert algorithm [18] to generate scale-free networks. We used Python's networkx package[1] to create a network with n_1=1024 nodes. We then perturb the adjacency matrix of the graph as follows to generate the second network:

$$A^{(2)} = P_N^\top A^{(1)} P_N + P_L, \tag{3}$$

where P_N is an $n_1 \times n_2$ binary matrix with at most one nonzero element in each row and column. If $n_1 = n_2$ then P_N is a permutation matrix. P_L is a symmetric matrix whose elements are chosen from $\{-1, 0, 1\}$ to perturb the network edges. After generating two random graphs, we construct feature vectors for the nodes by computing the following four node centrality measuresdegree, betweenness, closeness, and PageRank centrality [19]. Given the node features, we then calculate the cost matrix C based on their pairwise normalized Euclidean distance and set the feature similarity to be $S = \mathbf{1} - C$.

2) Real Data: We have used three real-world datasets from [5] for our experiments. In addition, we also consider the friendship network data from Deezer [20], a music streaming service, and Facebook [21]. For each friendship network, we generate its corresponding paired network by randomly sampling a subgraph using the random walk sampling approach to preserve its topological structure [22]. A summary of the datasets is presented in Table-I.

B. Experimental Setup

We consider both IsoRank [4] and FINAL [5] as baseline methods to assess the relative performance of OPTANE. For IsoRank and FINAL, we tuned their hyperparameter $\alpha \in \{0, 0.1, 0.15, \dots, 0.9, 1\}$, whereas for OPTANE, we performed a grid search for $\mu, \alpha \in \{1, 2, 10, 100\}$ and reported the results with highest accuracy. As the outputs of all three algorithms are non-binary, following the approach used in previous papers, we apply the heuristic greedy matching algorithm in [14], [5] to create a binary-valued alignment matrix T before computing the accuracy.

C. Experimental Results

For synthetic data, we investigated the effect of perturbing the nodes and link structure of the network using the approach given in Eqn. (3). First, we perturbed the nodes by randomly deleting 1%, 2%, and 5% of the nodes from the first graph. Second, we consider perturbing the structure by randomly adding or deleting 1%, 2%, or 5% of the links. We repeated the procedure 15 times to generate 15 different network pairs for our alignment experiments. The accuracies of the different methods are shown in Figure-V-B. The results suggest that OPTANE significantly outperformed the baseline methods especially when the amount of perturbed nodes and links are large. This demonstrates the robustness of OPTANE compared to the existing methods.

Figure-2 depicts the performance of OPTANE and other baseline algorithms on 5 real-world data sets. The results suggest that OPTANE outperforms IsoRank [4] in 4 of the 5 data sets and FINAL [5] in 3 of the 5 data sets.

D. Performance Comparison on Low Degree Nodes

We also examine how well the methods align low degree nodes in two networks. For this experiment, we restrict the analysis to nodes with degree less than 15. Since there are two networks, the low degree nodes here refer to those that belong to the network with smaller size. We calculated the percentage of correct prediction for nodes pair with degrees between 1 and 14. Figure-3 summarized the results for the Deezer real-world dataset. The gap in accuracy between OPTANE and the baselines increases as node degree decreases, which shows the robustness of the framework in terms of aligning low degree nodes.

VI. CONCLUSIONS

This paper presents an OT approach for network alignment problem. We demonstrated the effectiveness of our proposed OPTANE framework on both synthetic and real-world data and showed that OPTANE outperformed two state of the art methods in terms of its robustness to perturbed nodes and links as well as its effectiveness at aligning low degree nodes.

ACKNOWLEDGMENT

This work is supported in part by the U.S. National Science Foundation under grant #IIS-1615612.

REFERENCES

[1] Y. Zhang, "Browser-oriented universal cross-site recommendation and explanation based on user browsing logs," in *Proceedings of the 8th ACM Conference on Recommender systems*. ACM, 2014, pp. 433–436.

[2] R. Singh, J. Xu, and B. Berger, "Pairwise global alignment of protein interaction networks by matching neighborhood topology," in *Annual International Conference on Research in Computational Molecular Biology*. Springer, 2007, pp. 16–31.

[3] S. A. Cook, "The complexity of theorem-proving procedures," in *Proceedings of the third annual ACM symposium on Theory of computing*. ACM, 1971, pp. 151–158.

[4] R. Singh, J. Xu, and B. Berger, "Global alignment of multiple protein interaction networks with application to functional orthology detection," *Proceedings of the National Academy of Sciences*, vol. 105, no. 35, pp. 12 763–12 768, 2008.

[5] S. Zhang and H. Tong, "Final: Fast attributed network alignment," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2016, pp. 1345–1354.

[6] G. Monge, "Mémoire sur la théorie des déblais et des remblais," *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

[1] https://networkx.github.io

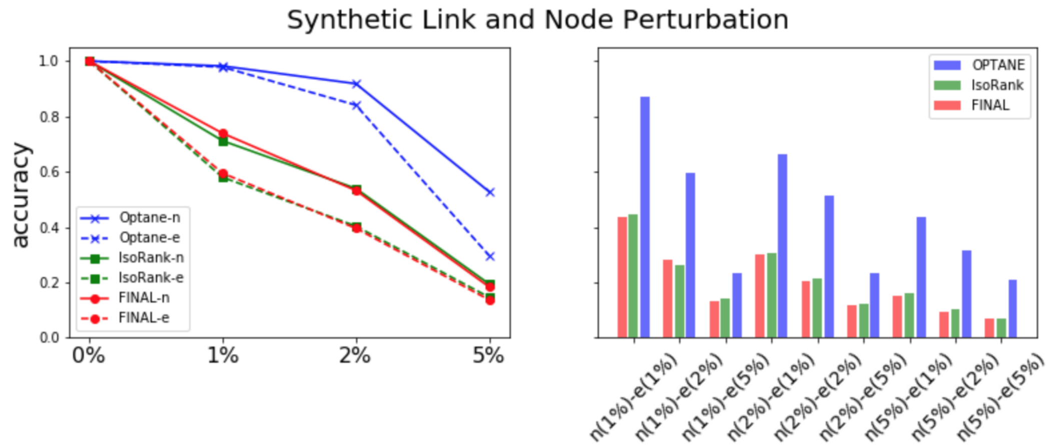

Figure 1. The node and edge perturbation. Left figure represents single perturbation scenarios -n used to show node perturbation and -e means edge perturbation for example Final-e is the line which represents accuracy of FINAL algorithm for 0% up to 5% edge perturbation. Right figure shows scenarios were with both node and edge perturbation, here $n(p\%) - e(q\%)$
means $p\%$ node perturbation, and $q\%$ edge perturbation are used to generate second network from first network

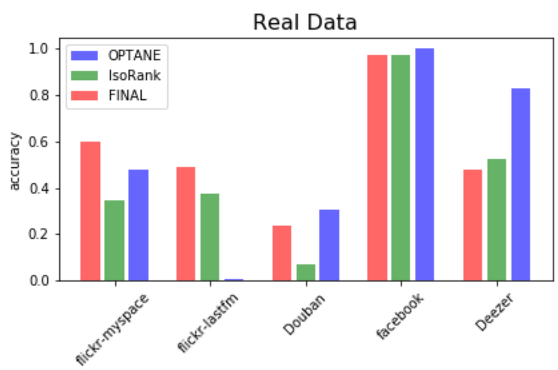

Figure 2. General accuracy of baseline algorithms on five real network datasets.

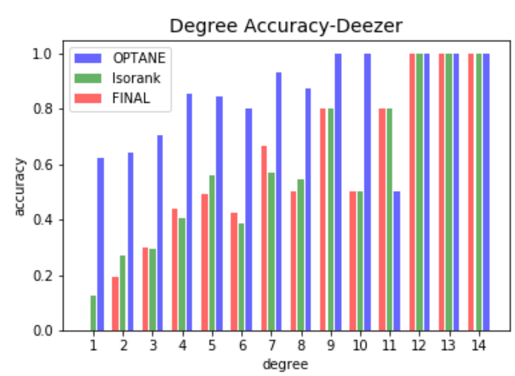

Figure 3. The percentage of correct prediction of alignment pair base on node in smaller network for Deezer music streaming service.

[7] G. Peyré, M. Cuturi *et al.*, "Computational optimal transport," Tech. Rep., 2017.

[8] R. Burkard, M. Dell'Amico, and S. Martello, "Assignment problems: Society for industrial and applied mathematics," 2009.

[9] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré, "Iterative bregman projections for regularized transportation problems," *SIAM Journal on Scientific Computing*, vol. 37, no. 2, pp. A1111–A1138, 2015.

[10] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang, "Algorithms for large, sparse network alignment problems," in *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*. IEEE, 2009, pp. 705–710.

[11] G. W. Klau, "A new graph-based method for pairwise global network alignment," *BMC bioinformatics*, vol. 10, no. 1, p. S59, 2009.

[12] D. Koutra, H. Tong, and D. Lubensky, "Big-align: Fast bipartite graph alignment," in *2013 IEEE 13th International Conference on Data Mining*. IEEE, 2013, pp. 389–398.

[13] J. T. Vogelstein, J. M. Conroy, L. J. Podrazik, S. G. Kratzer, D. E. Fishkind, R. J. Vogelstein, and C. E. Priebe, "Fast inexact graph matching with applications in statistical connectomics," *CoRR, abs/1112.5507*, 2011.

[14] G. Kollias, S. Mohammadi, and A. Grama, "Network similarity decomposition (nsd): A fast and scalable approach to network alignment," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 12, pp. 2232–2243, 2012.

[15] C.-S. Liao, K. Lu, M. Baym, R. Singh, and B. Berger, "Isorankn: spectral methods for global alignment of multiple protein networks," *Bioinformatics*, vol. 25, no. 12, pp. i253–i258, 2009.

[16] L. Kantorovitch, "On the translocation of masses," *Management Science*, vol. 5, no. 1, pp. 1–4, 1958.

[17] J. Renegar, "A polynomial-time algorithm, based on newton's method, for linear programming," *Mathematical Programming*, vol. 40, no. 1-3, pp. 59–93, 1988.

[18] A.-L. Barabási and R. Albert, "Emergence of scaling in random networks," *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[19] M. E. Newman, "Ego-centered networks and the ripple effect," *Social Networks*, vol. 25, no. 1, pp. 83–95, 2003.

[20] B. Rozemberczki, R. Davies, R. Sarkar, and C. Sutton, "Gemsec: graph embedding with self clustering," *arXiv preprint arXiv:1802.03997*, 2018.

[21] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, 2012, pp. 539–547.

[22] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2006, pp. 631–636.