From homogeneous to heterogeneous network alignment

Shawn Gu
Department of Computer Science and Engineering,
University of Notre Dame
Notre Dame, Indiana

John Johnson
Department of Computer Science and Engineering,
University of Notre Dame
Notre Dame, Indiana

Fazle E. Faisal
Department of Computer Science and Engineering,
University of Notre Dame
Notre Dame, Indiana

Tijana Milenković*
Department of Computer Science and Engineering,
University of Notre Dame
Notre Dame, Indiana
tmilenko@nd.edu

ABSTRACT

Network alignment (NA) aims to find a node mapping between compared networks that uncovers highly similar (i.e., conserved) network regions. Existing NA methods are homogeneous, i.e., they can deal only with networks containing nodes and edges of one type. However, due to increasing amounts of heterogeneous real-world network data with nodes and edges of different types, we extend three recent state-of-the-art homogeneous NA methods, WAVE, MAGNA++, and SANA, to allow for heterogeneous NA for the first time. We introduce several algorithmic novelties to achieve this. Namely, these existing NA methods compute homogeneous graphlet-based node similarities and then find high-scoring alignments with respect to these similarities, while simultaneously maximizing the number of conserved edges. Instead, we generalize homogeneous graphlets to their heterogeneous counterparts, which we then use to develop a new measure of heterogeneous node similarity. Also, we generalize S^3, a state-of-the-art measure of edge conservation for homogeneous NA, to its heterogeneous counterpart. Then, we find high-scoring alignments with respect to our heterogeneous node similarity and edge conservation measures. In evaluations on synthetic networks and real-world biological network data, our proposed heterogeneous NA methods lead to higher-quality alignments and better robustness to noise in the data than their homogeneous counterparts.

Note: This is a short, work-in-progress submission for a poster presentation at the HeteroNAM 2018 workshop that is not considered archival for resubmission purposes. The full paper version of this work is currently under review in a journal. After the full paper is published, the software implementing this work will be made publicly available. In the meanwhile, the software is available via email request.

1 INTRODUCTION

Traditional research about biological networks has focused on analyzing homogeneous networks containing a single node type and single edge type. For example, in protein-protein interaction (PPI) networks, nodes are proteins and edges are their interactions. Biological networks are important to study because they give a more complete picture of cellular functioning compared to looking at individual protein sequences: proteins carry out biological processes by interacting with each other, which is exactly what biological networks model. However, biological network data of different types are becoming available due to advancements of biotechnologies for their collection [4]. Integrating such data into a heterogeneous network, where nodes, edges, or both can be of different types (colors), and analyzing the heterogeneous data can lead to deeper biological insights compared to traditional analyses of homogeneous data (single node type and single edge type) [8]. For example, data have been collected on how proteins are related to diseases, and how drugs interact with proteins. When creating a heterogeneous network from these data, nodes would be proteins, diseases, or drugs, and edges would be protein-protein, disease-disease, drug-drug, protein-disease, protein-drug, or disease-drug interactions. Of course, there exist other entities and their associations that can be used as additional node and edge types.

Despite the potential improvements that heterogeneous network analysis can provide, many analyses only deal with homogeneous networks. This includes the task of network alignment (NA, described below). Therefore, if we want to study heterogeneous network data in this context, advances are needed to allow for heterogeneous NA (HNA). We first describe NA and its importance, and then introduce our modifications that allow three recent and thus state-of-the-art NA methods, WAVE [15], MAGNA++ [16], and SANA [9], to deal with heterogeneous networks. That is, our work enables HNA for the first time.

Biological NA aims to find a node mapping between the compared networks that uncovers regions of high similarity [10]. This allows for the transfer of functional knowledge between the similar (i.e., aligned) network regions. For example, if we align the PPI network of baker's yeast, a well-studied species, to the PPI network of human, a poorly-studied species, we can infer the function of human proteins based the function of their aligned partners in the yeast network. One important application of NA is the study of human aging, which is difficult to do experimentally due to long life span and ethical constraints. So, the current knowledge about human aging has been obtained mostly via computational analyses, namely genomic sequence alignment-based transfer of aging-related knowledge from well-studied model species to poorly-studied human. Recently we demonstrated that biological NA can be used for this same purpose, thus further deepening our currently limited knowledge about human aging [7].

*To whom correspondence should be addressed

2 METHODS

NA is computationally intractable, i.e., NP-hard [5]. So, heuristic methods need to be sought. These methods typically consist of two algorithmic parts. First, they measure the topological similarity between nodes from the networks being aligned. Then, they use an alignment strategy to quickly identify alignments that maximize the total topological similarity over all aligned nodes (or node conservation, NC, described below), and the amount of conserved edges (or edge conservation, EC, described below). As a proof of concept that HNA is worth studying, we modify the existing homogeneous WAVE, MAGNA++, and SANA methods to allow them to align heterogeneous networks. We hope that these extensions can improve alignment quality compared to their current homogeneous versions. We describe the modifications below.

The idea behind NC is as follows. Intuitively, two nodes from different networks are topologically similar if their extended neighborhoods are similar. Homogeneous NC of the above three methods quantifies this idea using graphlets. Graphlets are small, connected, induced subgraphs of a network (a path, triangle, square, etc.) [12]. For a given node, the graphlet-based NC counts how many times the node participates in each graphlet with up to n nodes (typically, n is 5). These counts form the node's *graphlet degree vector (GDV)*, which contains information on how the neighborhood of the node looks [11]. Then, GDVs of different nodes can be compared to quantify how similar the nodes are.

When we have a heterogeneous network, we can modify the above notion of similarity; now, two nodes from different networks are topologically similar if they are of the same color and if their extended neighborhoods are of similar color and structure. To capture this idea, we generalize current homogeneous graphlets to their heterogeneous (colored) counterparts. Given a homogeneous graphlet, its colored versions simply involve labeling each node with one of the possible colors. Then, as before, for a given node, the now-heterogeneous graphlet-based NC counts how many times the node participates in each colored graphlet. These counts form the node's *heterogeneous GDV*, the generalization of *homogeneous GDV*. These heterogeneous GDVs can be compared to quantify how similar the nodes are.

The idea behind EC is as follows. Aligning similar nodes is not enough to produce a good alignment. In addition, edges should be conserved. That is, if two nodes are connected in one network, then the nodes they are aligned to should be connected in the other. Homogeneous EC quantifies this with a measure called S^3 [13], scoring high if many edges are conserved; by optimizing this measure, NA algorithms can produce alignments with many conserved edges. With heterogeneous networks, we introduce heterogeneous EC, which considers the colors of the corresponding nodes to determine the degree to which the edges are conserved, meaning that matching colors is preferred. So, we modify S^3 to account for colors of the nodes involved in the conserved edges, resulting in our new heterogeneous EC.

WAVE. WAVE takes as input two networks and similarities between all pairs of nodes from the two networks. Then, it uses a "seed-and-extend" algorithm to align the networks. First two highly similar nodes are aligned, i.e., seeded. Then, the seed's neighboring nodes (or simply neighbors) that are similar are aligned, the seed's neighbor's neighbors that are similar are aligned, and so on. This step of extending around the seed and exploring the seed's neighbors is intended to improve both NC and EC of the resulting alignment. The extension step continues until all nodes in the smaller of the two compared networks are aligned (formally, until a one-to-one node mapping between the two networks is produced). So, to modify WAVE to be heterogeneous, we simply give it node similarities based on heterogeneous graphlets; we want to favor aligning nodes of the same color, which is captured by our heterogeneous graphlet similarities. Note that because the algorithm looks at neighbors of the seed, WAVE only implicitly optimizes EC, so there is no opportunity to use our heterogeneous EC as a parameter.

MAGNA++ and SANA. MAGNA++ and SANA take as input two networks and similarities between all pairs of nodes from the two networks. However, these methods use search algorithms to find an alignment. That is, instead of aligning networks node by node as WAVE does, MAGNA++ and SANA explore the space of possible alignments and find the highest scoring one with respect to some measure of NC and EC. In the homogeneous case, both methods perform well when optimizing a combination of a graphlet-based NC measure and the S^3 EC measure. So, to modify MAGNA++ and SANA to be heterogeneous, we modify the methods to optimize our heterogeneous graphlet-based NC and our heterogeneous S^3 measure. Note that MAGNA++ uses a genetic algorithm as its search strategy, while SANA uses simulated annealing.

3 RESULTS AND DISCUSSION

We evaluate the three HNA methods versus their homogeneous versions in three tests. First, we align synthetic networks, which originate from three different network models (geometric/GEO, scale-free/SF, and Erdős-Rényi/ER) and which have up to four artificially imposed node colors, to their noisy counterparts (defined below). Second, we align homogeneous (single-node-type) PPI networks, which have up to four node colors imposed according to proteins' involvement in a combination of aging [3], [6], cancer [2], and Alzheimer disease [14], to their noisy counterparts. Here, node colors originate from gene expression (Expr) [3] or sequence (Seq) [6] analyses. Also, here, we consider each of three types of PPIs: only affinity capture coupled to mass spectrometry (APMS), only two-hybrid (Y2H), and both combined (APMS+Y2H). Third, we align heterogeneous (two-node-type) biological networks to their noisy counterparts. Here, the two node types/colors correspond to proteins and their Gene Ontology (GO) terms [1], and edges exist between proteins (where we again use each of the three types of PPIs), between GO terms, and between proteins and GO terms. Note that we do not test MAGNA++ on the large PPI and protein-GO networks since MAGNA++ is not as scalable on such large data as the other two methods.

A noisy counterpart is the original network with x% of its edges rewired, where we vary x from 0 to 50 in increments of 10. Since only edges are changed, we know which nodes should be mapped to which. We can use this true node mapping to accurately evaluate our methods; a good method should have high node correctness, which is the percentage of node pairs from the given alignment that are present in the true node mapping. Varying the percentage

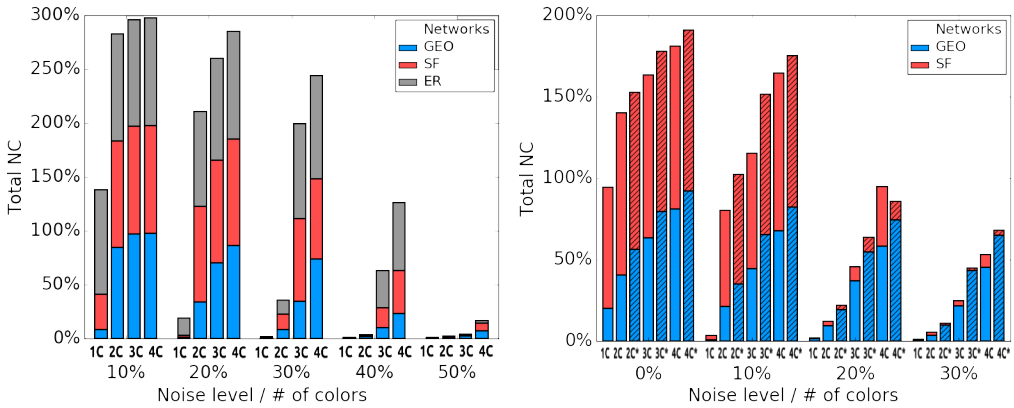

Figure 1: Node correctness (totaled over all analyzed networks in the given test) vs. noise level/number of colors for synthetic networks using (a) WAVE and (b) MAGNA++. Note that "1C" corresponds to homogeneous WAVE/MAGNA++, "2C" corresponds to heterogeneous WAVE/MAGNA++ with two colors, "3C" corresponds to heterogeneous WAVE/MAGNA++ with three colors, and "4C" corresponds to heterogeneous WAVE/MAGNA++ with four colors. In panel (b), solid bars represent using hetNC-homEC, while dashed bars (also marked with *) represent using hetNC-hetEC.

of edges we rewire also lets us test how robust our methods are to noise, which is useful since biological networks are usually noisy.

We expect that for a given noise level, HNA (2 or more colors) should improve alignment quality over homogeneous NA (1 color). Also, we expect that the more colors are used, the better the alignment quality should be, since we are using more information in the process of producing the alignment. In addition, we predict that using more colors will make the given method more robust to noise, meaning that we should see a slower decrease in alignment quality as noise increases, compared to using fewer colors. Note, however, that alignment quality should be low at the highest noise levels regardless of how many colors we use, since we are essentially aligning two networks with almost random topologies compared to each other.

In Figure 1 we report a subset of these results, specifically those for WAVE and MAGNA++ on synthetic networks; these methods' results for PPI and protein-GO networks, and results of SANA for all three networks sets are qualitatively similar, and the complete set of results will be made available in the full journal version of our paper upon its publication. In particular, for synthetic networks, across all evaluation tests, using 4 colors performs the best 100% of the time, using 3 colors performs the best 10% of the time, using 2 colors performs the best 3% of the time, and using 1 color performs the best 0% of the time. For PPI networks, across all evaluation tests, using 4 colors performs the best 95% of the time, using 3 colors performs the best 5% of the time, using 2 colors performs the best 0% of the time, and using 1 color performs the best 0% of the time. For protein-GO networks, across all evaluation tests, using 2 colors performs the best 100% of the time, and using 1 color performs the best 0% of the time. Note that ties can occur for different numbers of colors. According to the above results, clearly, HNA improves upon homogeneous NA. Also, the more colors are considered, the better the alignment quality. Finally, as shown in Figure 1, the more colors, the more robust the alignments are to noise. In summary, all of our expectations (see above) have been met.

All of the above results are when modifying homogeneous NC into heterogeneous NC within the existing methods, while still using homogeneous EC (call this scenario hetNC-homEC). When we also modify homogeneous EC into heterogeneous EC (call this scenario hetNC-hetEC), we observe that hetNC-hetEC performs as least as good or better than hetNC-homEC 90%, 100%, and 93% of the time for synthetic, PPI, and protein-GO networks, respectively. That is, each of heterogeneous NC and heterogeneous EC contributes to improved alignment quality. In Figure 1b we report raw results for MAGNA++ on synthetic networks, comparing hetNC-homEC (solid bars) to hetNC-hetEC (dashed bars). Clearly, in almost all cases, a given dashed bar is higher than the corresponding solid bar.

In summary, we modify WAVE, MAGNA++, and SANA to align heterogeneous networks by extending the existing notions of NC and EC to their heterogeneous counterparts. Specifically, we extend homogeneous graphlets to their heterogeneous counterparts, and homogeneous S^3 to heterogeneous S^3. We evaluate our methods by aligning synthetic, PPI, and protein-GO networks to their noisy counterparts. We show that using more colors leads to better alignments, and that using heterogeneous EC over homogeneous EC (while still using heterogeneous NC) also leads to increases in alignment quality. Because of these improvements of HNA over homogeneous NA, it is useful to continue to study and compare heterogeneous networks.

REFERENCES

[1] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. 2000. Gene Ontology: tool for the unification of biology. *Nature genetics* 25, 1 (2000), 25.

[2] Sally Bamford, E Dawson, Simon Forbes, Jody Clements, Roger Pettett, Ahmet Dogan, A Flanagan, Jon Teague, P Andrew Futreal, Michael R Stratton, et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *British journal of cancer* 91, 2 (2004), 355.

[3] Nicole C Berchtold, David H Cribbs, Paul D Coleman, Joseph Rogers, Elizabeth Head, Ronald Kim, Tom Beach, Carol Miller, Juan Troncoso, John Q Trojanowski, et al. 2008. Gene expression changes in the course of normal brain aging are sexually dimorphic. *Proceedings of the National Academy of Sciences* 105, 40 (2008), 15605–15610.

[4] Bobby-Joe Breitkreutz, Chris Stark, Teresa Reguly, Lorrie Boucher, Ashton Breitkreutz, Michael Livstone, Rose Oughtred, Daniel H. Lackner, JÃijrg BÃďhler, Valerie Wood, Kara Dolinski, and Mike Tyers. 2008. The BioGRID Interaction Database: 2008 update. *Nucleic Acids Research* 36, suppl_1 (2008), D637–D640. https://doi.org/10.1093/nar/gkm1001

[5] Stephen A Cook. 1971. The complexity of theorem-proving procedures. In *Proceedings of the third annual ACM symposium on Theory of computing*. ACM, 151–158.

[6] João Pedro de Magalhães. 2009. Aging research in the post-genome era: New technologies for an old problem. *Redox Metabolism and Longevity Relationships in Animals and Plants. Taylor and Francis, New York and Abingdon* (2009), 99–115.

[7] Fazle Elahi Faisal, Han Zhao, and Tijana Milenković. 2015. Global network alignment in the context of aging. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 12, 1 (2015), 40–52.

[8] Vladimir Gligorijević and Nataša Pržulj. 2015. Methods for biological data integration: perspectives and challenges. *Journal of the Royal Society Interface* 12, 112 (2015), 20150571.

[9] Nil Mamano and Wayne B. Hayes. 2017. SANA: simulated annealing far outperforms many other search algorithms for biological network alignment. *Bioinformatics* 33, 14 (2017), 2156–2164. https://doi.org/10.1093/bioinformatics/btx090

[10] Lei Meng, Aaron Striegel, and Tijana Milenković. 2016. Local versus global biological network alignment. *Bioinformatics* 32, 20 (2016), 3155–3164.

[11] Tijana Milenković and Nataša Pržulj. 2008. Uncovering Biological Network Function via Graphlet Degree Signatures. *Cancer Informatics* 6 (2008). https://doi.org/10.4137/cin.s680

[12] Natasa Pržulj, Derek G Corneil, and Igor Jurisica. 2004. Modeling interactome: scale-free or geometric? *Bioinformatics* 20, 18 (2004), 3508–3515.

[13] Vikram Saraph and Tijana Milenković. 2014. MAGNA: maximizing accuracy in global network alignment. *Bioinformatics* 30, 20 (2014), 2931–2940.

[14] Julie E Simpson, Paul G Ince, Pamela J Shaw, Paul R Heath, Rohini Raman, Claire J Garwood, Catherine Gelsthorpe, Lynne Baxter, Gillian Forster, Fiona E Matthews, et al. 2011. Microarray analysis of the astrocyte transcriptome in the aging brain: relationship to Alzheimer's pathology and APOE genotype. *Neurobiology of aging* 32, 10 (2011), 1795–1807.

[15] Yihan Sun, Joseph Crawford, Jie Tang, and Tijana Milenković. 2015. Simultaneous Optimization of both Node and Edge Conservation in Network Alignment via WAVE. *Lecture Notes in Computer Science Algorithms in Bioinformatics* (2015), 16–39. https://doi.org/10.1007/978-3-662-48221-6_2

[16] V. Vijayan, V. Saraph, and T. Milenković. 2015. MAGNA++: Maximizing Accuracy in Global Network Alignment via both node and edge conservation. *Bioinformatics* 31, 14 (2015), 2409–2411. https://doi.org/10.1093/bioinformatics/btv161