AFRL-AFOSR-VA-TR-2019-0356

Efficient comparison of multiple complex networks

Tijana Milenkovic
UNIVERSITY OF NOTRE DAME DU LAC

09/06/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
03-09-2019	Final	Jul 2016 - Jun 2019

4. TITLE AND SUBTITLE
(YIP) Efficient Comparison of Multiple Complex Networks

5a. CONTRACT NUMBER
N/A

5b. GRANT NUMBER
FA9550-16-1-0147

5c. PROGRAM ELEMENT NUMBER
N/A

6. AUTHOR(S)
Milenkovic, Tijana

5d. PROJECT NUMBER
N/A

5e. TASK NUMBER
N/A

5f. WORK UNIT NUMBER
N/A

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Notre Dame
940 Grace Hall
Notre Dame, IN 46556-5708

8. PERFORMING ORGANIZATION REPORT NUMBER
N/A

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFOSR
875 North Randolph Street
Suite 325, Room 3112
Arlington, VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
N/A

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for Public Release; distribution is Unlimited

13. SUPPLEMENTARY NOTES
N/A

14. ABSTRACT
Network alignment (NA), one of the most popular network science/mining tasks, aims to compare networks corresponding to different systems in order to identify network regions of the systems' (dis)similarities, thus allowing for learning something about a poorly understood system from a well understood system based on their aligned network regions. As such, NA has applications in a variety of domains, including computational biology, chemoinformatics, neuroscience, computational linguistics, artificial intelligence, computer vision, and web mining. Since complexity theory dictates that the problem of NA is computationally hard, this project introduced novel computationally efficient yet accurate heuristic NA approaches, such as those for alignment of multiple networks (as opposed to traditional pairwise NA), dynamic networks (as opposed to traditional static NA), or heterogeneous networks (as opposed to traditional homogeneous NA). The project resulted in 10 published or submitted papers and 16 conference presentations of the project results. It supported eight researchers (the principal investigator, a postdoctoral researcher, four Ph.D. students, and two undergraduate students).

15. SUBJECT TERMS
Network alignment, network comparison, topological similarity, node/edge conservation, network/node embedding, graphlets/subgraphs, subgraph isomorphism, dynamic networks, heterogeneous networks

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Tijana Milenkovic
U	U	U	UU	13	19b. TELEPHONE NUMBER *(Include area code)* (574) 631-8975

INSTRUCTIONS FOR COMPLETING SF 298

1. REPORT DATE. Full publication date, including day, month, if available. Must cite at least the year and be Year 2000 compliant, e.g. 30-06-1998; xx-06-1998; xx-xx-1998.

2. REPORT TYPE. State the type of report, such as final, technical, interim, memorandum, master's thesis, progress, quarterly, research, special, group study, etc.

3. DATE COVERED. Indicate the time during which the work was performed and the report was written, e.g., Jun 1997 - Jun 1998; 1-10 Jun 1996; May - Nov 1998; Nov 1998.

4. TITLE. Enter title and subtitle with volume number and part number, if applicable. On classified documents, enter the title classification in parentheses.

5a. CONTRACT NUMBER. Enter all contract numbers as they appear in the report, e.g. F33315-86-C-5169.

5b. GRANT NUMBER. Enter all grant numbers as they appear in the report. e.g. AFOSR-82-1234.

5c. PROGRAM ELEMENT NUMBER. Enter all program element numbers as they appear in the report, e.g. 61101A.

5e. TASK NUMBER. Enter all task numbers as they appear in the report, e.g. 05; RF0330201; T4112.

5f. WORK UNIT NUMBER. Enter all work unit numbers as they appear in the report, e.g. 001; AFAPL30480105.

6. AUTHOR(S). Enter name(s) of person(s) responsible for writing the report, performing the research, or credited with the content of the report. The form of entry is the last name, first name, middle initial, and additional qualifiers separated by commas, e.g. Smith, Richard, J, Jr.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES). Self-explanatory.

8. PERFORMING ORGANIZATION REPORT NUMBER. Enter all unique alphanumeric report numbers assigned by the performing organization, e.g. BRL-1234; AFWL-TR-85-4017-Vol-21-PT-2.

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES). Enter the name and address of the organization(s) financially responsible for and monitoring the work.

10. SPONSOR/MONITOR'S ACRONYM(S). Enter, if available, e.g. BRL, ARDEC, NADC.

11. SPONSOR/MONITOR'S REPORT NUMBER(S). Enter report number as assigned by the sponsoring/ monitoring agency, if available, e.g. BRL-TR-829; -215.

12. DISTRIBUTION/AVAILABILITY STATEMENT. Use agency-mandated availability statements to indicate the public availability or distribution limitations of the report. If additional limitations/ restrictions or special markings are indicated, follow agency authorization procedures, e.g. RD/FRD, PROPIN, ITAR, etc. Include copyright information.

13. SUPPLEMENTARY NOTES. Enter information not included elsewhere such as: prepared in cooperation with; translation of; report supersedes; old edition number, etc.

14. ABSTRACT. A brief (approximately 200 words) factual summary of the most significant information.

15. SUBJECT TERMS. Key words or phrases identifying major concepts in the report.

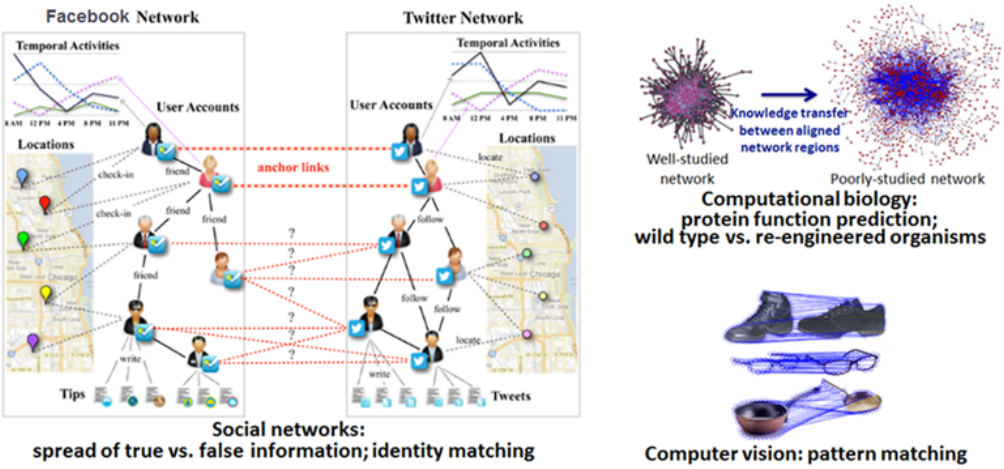
16. SECURITY CLASSIFICATION. Enter security classification in accordance with security classification regulations, e.g. U, C, S, etc. If this form contains classified information, stamp classification level on the top and bottom of this page.

17. LIMITATION OF ABSTRACT. This block must be completed to assign a distribution limitation to the abstract. Enter UU (Unclassified Unlimited) or SAR (Same as Report). An entry in this block is necessary if the abstract is to be limited.

Final project report for award AFOSR YIP FA9550-16-1-0147

1 Research aims

Network alignment (NA), one of the most popular network science/mining tasks, aims to compare networks corresponding to different systems in order to identify network regions of the systems' (dis)similarities, thus allowing for learning something about a poorly understood system from a well understood system based on their aligned network regions. As such, NA has applications in a variety of domains, including computational biology, chemoinformatics, neuroscience, computational linguistics, artificial intelligence, computer vision, and web mining (see **Figure 1** for some specific illustrations).



Applications:

- Social networks:
  - The spread of true vs. false information
  - Entity (user) matching across different online platforms (cybersecurity/privacy implications)
- Computational biology:
  - From genomic sequence to biological network comparison
  - Wild type vs. genetically re-engineered organisms (e.g., defense against bioweapons)
  - Drug-sensitive vs. drug-resistant organisms
- Computational neuroscience:
  - Functional brain networks of healthy vs. disease-affected people (e.g., PTSD)
- Computational linguistics:
  - Networks of different languages → machine translation
- …

Figure 1: Practical applications of NA (many of which have **not yet** been explored). Note that the social network figure on the left has been adapted from https://www.springer.com/gp/book/9783319562117 .

Since complexity theory dictates that the problem of NA is computationally intractable (formally, NP-hard), this project introduced novel computationally efficient yet accurate heuristic NA approaches. As such, with its state-of-the-art novelties, this project has significantly advanced the field of NA. Specifically, while the majority of the existing NA efforts have focused on pairwise NA (PNA), i.e., NA of two networks, this project explored multiple NA (MNA), i.e., NA of more than two networks. Also, while all existing NA efforts focused on NA of static networks, this project proposed the first ever methods for NA of dynamic networks. Similarly, while all existing NA efforts focused on NA of homogeneous (single node/edge type) networks, this project proposed the first ever methods for NA of heterogeneous (multiple node/edge type) networks. Additional algorithmic developments for NA were proposed as well. In total, this project spanned four novel computational aims for efficient yet accurate NA, which were completed successfully:

1. **Efficient shift from PNA to MNA:** extend the theory behind a recent state-of-the-art PNA approach, called MAGNA++, into a novel MNA approach, called multiMAGNA++.
2. **Including network "geometricity" into NA:** since some real-world networks have a "geometric" aspect, in the sense that an order can be imposed on the nodes of the given network, allow for including the order information into the network alignment process.
3. **Further algorithmic NA developments:** develop new computational approaches that further improve network alignment quality.
4. **NA of dynamic networks:** allow for aligning temporal (dynamic, evolving) networks, unlike all existing PNA or MNA approaches, which can deal only with static networks.

2 Research accomplishments (covering the entire 2016-2019 project period)

The project resulted in a total of **10 research papers**: six published journal papers, one published conference paper, and three submitted papers that are currently under review, as follows:

1. Vipin Vijayan and Tijana Milenkovic (2017), **Multiple network alignment via multiMAGNA++**, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5): 1669-1682 (journal impact factor: 2.428).
2. Fazle E. Faisal, Khalique Newaz, Julie L. Chaney, Jun Li, Scott J. Emrich, Patricia L. Clark, and Tijana Milenkovic (2017), **GRAFENE: Graphlet-based alignment-free network approach integrates 3D structural and sequence (residue order) data to improve protein structural comparison**, *Nature Scientific Reports*, 7, Article number: 14890 (journal impact factor: 4.525).
3. Vipin Vijayan, Dominic Critchlow, and Tijana Milenkovic (2017), **Alignment of dynamic networks**, *Bioinformatics*, 33(14): i180-i189 (journal impact factor: 4.531).
4. Vipin Vijayan and Tijana Milenkovic (2018), **Aligning dynamic networks with DynaWAVE**, *Bioinformatics*, 34(10): 1795–1798 (journal impact factor: 4.531).
5. Shawn Gu, John Johnson, Fazle E. Faisal, and Tijana Milenkovic (2018), **From homogeneous to heterogeneous network alignment via colored graphlets**, *Nature Scientific Reports,* 8, Article number: 12524 (journal impact factor: 4.525).
6. Shawn Gu and Tijana Milenkovic (2018), **Graphlets versus node2vec and struc2vec in the task of network alignment**, In Proceedings of the 14th International Workshop on Mining and Learning with Graphs (MLG) at the 24th ACM SIGKDD 2018 Conference on Knowledge Discovery & Data Mining (KDD), London, UK, August 19-23, 2018.
7. David Aparicio, Pedro Ribeiro, Tijana Milenkovic, and Fernando Silva (2019), **Temporal network alignment via GoT-WAVE**, *Bioinformatics*, doi: 10.1093/bioinformatics/btz119 (journal impact factor: 4.531).
8. Vipin Vijayan, Eric Krebs, and Tijana Milenkovic (2019), **Pairwise versus multiple network alignment**, under review. Also, arXiv:1709.04564 [q-bio.MN].
9. Shawn Gu and Tijana Milenkovic (2019), **Data-driven network alignment**, under review. Also, arXiv:1902.03277 [q-bio.MN].
10. Shikang Liu, Fatemeh Vahedian, David Hachen, Omar Lizardo, Christian Poellabauer, Aaron Striegel, and Tijana Milenkovic (2019), **Heterogeneous network approach to predict individuals' mental health**, under review. Also, arXiv:1906.04346 [cs.SI].

The project resulted in a total of **16 conference presentations**: 11 oral plus nine poster conference presentations, at the following venues:

1. International Workshop on Data Mining in Bioinformatics (BIOKDD) at ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD), San Francisco, CA, USA, August 13-17, 2016 (oral presentation).
2. The Main Track at International School and Conference on Network Science (NetSci), Indianapolis, IN, June 19-23, 2017 (oral as well as poster presentation).
3. Great Lakes Bioinformatics Conference (GLBIO), Chicago, IL, May 15-17, 2017 (poster presentation).
4. Network Medicine (NetMed) Satellite Meeting at International School and Conference on Network Science (NetSci), Indianapolis, IN, June 19 - 23, 2017 (oral presentation).
5. The Main Track at International School and Conference on Network Science (NetSci), Indianapolis, IN, June 19 - 23, 2017 (poster presentation).
6. Network Models in Cellular Regulation (NetSciReg) Satellite Meeting at International School and Conference on Network Science (NetSci), Indianapolis, IN, June 19 - 23, 2017 (oral presentation).
7. 3D-SIG Community of Special Interest (COSI) at International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB), Prague, Czech Republic, July 21-25, 2017 (poster presentation).
8. Network Biology (NetBio) Community of Special Interest (COSI) International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB), Prague, Czech Republic, July 21-25, 2017 (oral as well as poster presentation).
9. International Workshop on Heterogeneous Network Analysis and Mining (HeteroNAM) in conjunction with International Conference on Web Search and Data Mining (WSDM), Los Angeles, CA, USA, Feb. 5-9, 2018 (oral presentation).
10. International School and Conference on Network Science (NetSci), Paris, France, June 11-15, 2018 (poster presentation).
11. Personalized Medicine in the era of Big Data (NetMed) Satellite in conjunction with International School and Conference on Network Science (NetSci), Paris, France, June 11-15, 2018 (poster presentation).
12. Machine Learning in Network Science (MLNS) Satellite in conjunction with International School and Conference on Network Science (NetSci), Paris, France, June 11-15, 2018 (oral presentation).
13. 3D-SIG Community of Special Interest (COSI) at International Conference on Intelligent Systems for Molecular Biology (ISMB), Chicago, IL, USA, July 6-10, 2018 (oral as well as poster presentation).
14. Network Biology (NetBio) COSI at International Conference on Intelligent Systems for Molecular Biology (ISMB), Chicago, IL, USA, July 6-10, 2018 (oral as well as poster presentation).
15. Workshop on Data Mining in Bioinformatics (BIOKDD) at the ACM SIGKDD 2018 Conference on Knowledge Discovery and Data Mining (KDD), London, UK, August 19-23, 2018 (oral presentation).

16. Network Biology (NetBio) COSI at International Conference on Intelligent Systems for Molecular Biology and European Conference on Computational Biology (ISMB/ECCB), Basel, Switzerland, July 21-24, 2019 (oral as well as poster presentation).

The project supported a total of **eight researchers**:

- The principal investigator, Dr. Milenkovic.
- A postdoc, Dr. Fatemeh Vahedian.
- Four Ph.D. students (Dr. Vipin Vijayan, Dr. Fazle Faisal, Khalique Newaz, and Shawn Gu), two of whom graduated.
- Two undergraduate researchers (Dominic Critchlow and Eric Krebs), both of whom graduated.

2.1 Aim 1: Efficient shift from PNA to MNA

We achieved two key accomplishments in this aim.

First, existing MNA methods aim to maximize total similarity over all aligned nodes (node conservation). Then, they evaluate alignment quality by measuring the amount of conserved edges, but only after the alignment is constructed. Directly optimizing edge conservation during alignment construction in addition to node conservation may result in superior alignments. Thus, we introduced a novel MNA approach called multiMAGNA++ that can achieve this. Indeed, multiMAGNA++ generally outperforms or is on par with the existing MNA methods (**Figure 2**), while often completing faster than the existing methods. That is, multiMAGNA++ scales well to larger network data and can be parallelized effectively. During method evaluation, we also introduced new MNA quality measures to allow for more complete alignment characterization as well as more fair MNA method comparison compared to using only the existing alignment quality measures. All code and data are available at: https://nd.edu/~cone/multiMAGNA++/ .

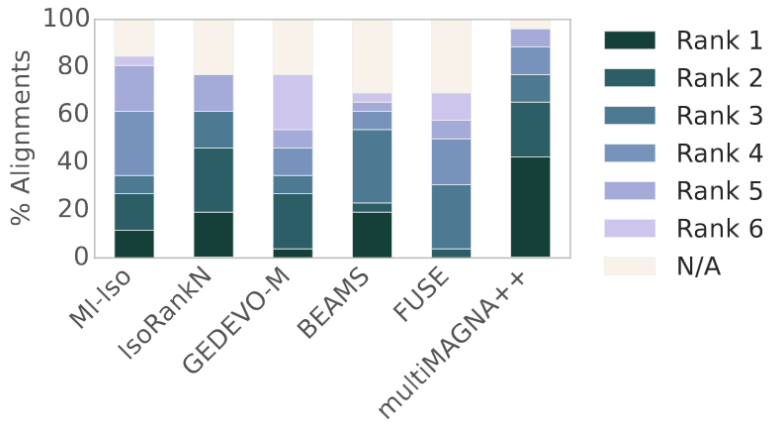


Figure 2: Ranking of multiMAGNA++ against five existing MNA methods (MI-Iso, IsoRankN, GEDEVO-M, BEAMS, and FUSE) across all five analyzed network sets with respect to all six analyzed measures of topological or functional alignment quality. The ranking of each method is expressed as a percentage of all evaluation tests (i.e., combinations of network sets and alignment quality measures) in which the given method is the best performing ("Rank 1"), the second best performing ("Rank 2"), etc. aligner of all considered methods. If an alignment score of a method is not statistically significant, the method is not ranked and is labelled as "N/A". Clearly, multiMAGNA++ is ranked the best in most of the evaluation tests, and it also has the fewest of non-significant alignment scores.

The above work was published as follows:
Vipin Vijayan and Tijana Milenkovic (2017), **Multiple network alignment via multiMAGNA++**, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(5): 1669-1682.

Second, recall that network alignment (NA) can be pairwise (PNA) and multiple (MNA). PNA produces aligned node pairs between two networks. MNA produces aligned node clusters between more than two networks. Recently, the focus (including our work on multiMAGNA++ in **Aim 1** of this project) has shifted from PNA to MNA, because MNA captures conserved regions between more networks than PNA (and MNA is thus considered to be more insightful), though at higher computational complexity. The issue is that, due to the different outputs of PNA and MNA, a PNA method is only compared to other PNA methods, and an MNA method is only compared to other MNA methods. Comparison of PNA against MNA must be done to evaluate whether MNA's higher complexity is justified by its higher accuracy. So, as a part of this project, in its year 2, we introduced a framework that allows for this. We compared PNA against MNA in both a pairwise (native to PNA) and multiple (native to MNA) manner. Shockingly, we found that PNA is more accurate and faster than MNA in both cases. This striking result will guide future research efforts in the NA field. Basically, we devised a strategy that can "simply" align all pairs of networks and integrate the resulting pairwise alignments into a multiple one, which results in both higher accuracy and lower running time compared to aligning all (multiple) networks at once. So, it is questionable whether MNA as traditionally defined will even be needed any more.

The above work has resulted in the following paper:
Vipin Vijayan, Eric Krebs, and Tijana Milenkovic (2019), **Pairwise versus multiple network alignment**, under review. Also, arXiv:1709.04564 [q-bio.MN].

2.2 Aim 2: Including network "geometricity" into NA

Recall that multiMAGNA++ from **Aim 1** optimizes both node and edge conservation during alignment construction, where the measure of node conservation is the total "topological similarity" over all aligned nodes, according to which two nodes from different networks are similar if their *graphlet-based* "topological signatures", i.e., the nodes' extended network neighborhoods, are similar; graphlets are small subgraphs, i.e., Lego-like building blocks, of a network. Given this, and given the existence of networks with known node order (for example, in protein structure networks that model spatial interactions between amino acids in proteins' 3-dimensional crystal structures, the order of nodes corresponds to the amino acids' positions in the protein sequences), in **Aim 2**, we proposed to generalize the established notion of graphlet-based topological signature of a node to its *ordered graphlet*-based counterpart. The goal of imposing node order onto a graphlet was to develop a more precise topological signature of a node (two nodes could have identical signatures with respect to regular (non-ordered) graphlets but different signatures with respect to ordered graphlets) and thus allow for more precise graphlet-based measure of node conservation to be used within our multiMAGNA++ approach from **Aim 1** or within other NA (and in general, network comparison) approaches.

In this context, we developed an approach based on the idea of ordered graphlets that allows for pairwise (rather than multiple) network comparison. Also, the approach allows for alignment-free (rather than alignment-based) network comparison. By alignment-free network comparison, we mean that the current approach "simply" aims to quantify the similarity between networks, without accounting for the mapping between their nodes, and without being able to identify the actual regions of similarities between the compared networks. In contrast, alignment-based network comparison (or simply NA) explicitly aims to find a node mapping that identifies similar (conserved) regions between the compared networks.

We applied this (pairwise and alignment-free) approach to the task of protein structural comparison. We chose this particular application because protein structure networks (PSNs) are a natural choice of real-world networks that contain node order (see above), and because this application directly fits another one of our funded projects, NIH award 1R01GM120733 (titled "Integrative computational framework for pattern mining in big -omics data: linking synonymous codon usage to protein biogenesis"). Specifically, initial protein structural comparisons were sequence-based. Since amino acids that are distant in the sequence can be close in the 3-dimensional (3D) structure, 3D contact approaches can complement sequence approaches. Traditional 3D contact approaches study 3D structures directly. Instead, 3D structures can be modeled as PSNs. Then, network approaches can compare proteins by comparing their PSNs. We hypothesized that network approaches may improve upon traditional 3D contact approaches. We could not use existing PSN approaches to test this, because: 1) They rely on naive measures of network topology. 2) They are not robust to PSN size. They cannot integrate 3) multiple PSN measures or 4) PSN data with sequence data, although this could help because the different data types capture complementary biological knowledge. We addressed these limitations by: 1) exploiting well-established graphlet measures via a new network approach, 2) introducing normalized graphlet measures to remove the bias of PSN size, 3) allowing for integrating multiple PSN measures, and 4) using ordered graphlets to combine the complementary PSN data and sequence (amino acid residue order) data. Our approach compares both synthetic networks and real-world PSNs more accurately and faster than existing network, 3D contact, or sequence approaches (**Figure 3** and **Table 1**). All code and data are available at https://nd.edu/~cone/PSN/.

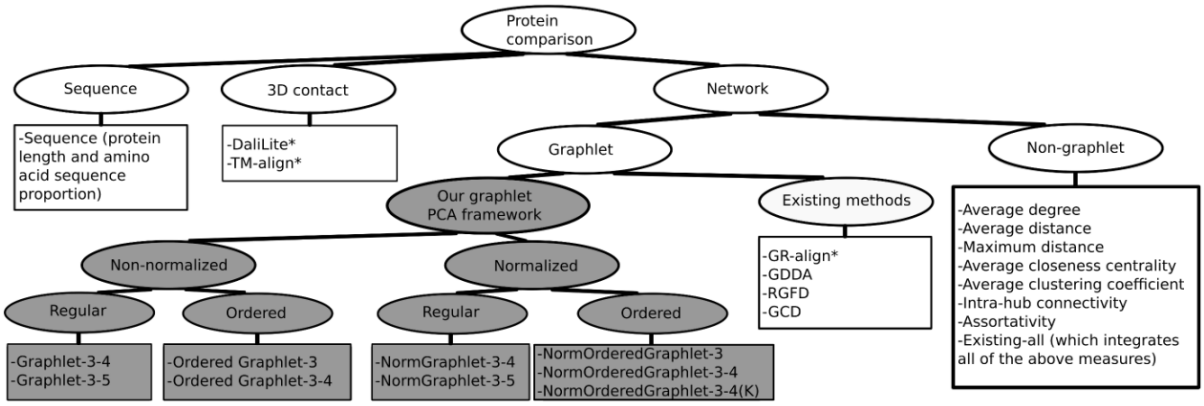


Figure 3: Categorization of the 24 protein structural comparison approaches (in squares) that we evaluated. Different versions of our new graphlet approach are colored in grey; all remaining approaches are existing ones. Alignment-based approaches are marked with *; all remaining approaches are alignment-free.

Approach	AUPR		AUROC		Running
	Rank	*p*-value	Rank	*p*-value	time (hrs)
Graphlet-3-4	8.38	9.42e-05	10.50	0.000147	0.43
Graphlet-3-5	9.00	4.81e-06	10.40	8.74e-05	0.49
OrderedGraphlet-3	7.15	0.00225	9.92	0.000692	0.38
OrderedGraphlet-3-4	7.31	0.00143	8.69	0.0018	2.39
NormGraphlet-3-4	7.77	3.57e-05	8.15	0.000156	0.44
NormGraphlet-3-5	8.15	5.04e-05	6.69	0.00124	0.51
NormOrderedGraphlet-3	10.50	4.33e-05	9.92	0.000135	0.39
NormOrderedGraphlet-3-4	4.31	0.000999	4.92	0.00127	2.41
NormOrderedGraphlet-3-4(K)	**1.69**	-	**2.08**	-	2.41
GDDA	17.30	6.16e-09	17.70	2.57e-08	0.54
RGFD	9.46	6.84e-06	9.85	1.39e-05	0.49
GCD	17.10	1.21e-09	17.10	1.51e-08	1.32
GR-align	8.31	0.00705	9.69	0.00423	9.49
Average degree	18.90	2.32e-10	16.20	2.02e-07	0.39
Average distance	15.40	9.54e-07	16.50	3.59e-06	0.48
Maximum distance	17.30	1.58e-09	16.90	4.95e-08	0.49
Average closeness centrality	18.50	2.18e-08	16.50	3.08e-07	0.48
Average clustering coefficient	16.80	5.01e-08	14.50	3.55e-07	0.56
Intra-hub connectivity	16.40	2.84e-08	15.10	1.14e-06	0.64
Assortativity	20.10	1.79e-08	19.20	1.48e-07	0.46
Existing-all	10.90	1.33e-06	10.00	3.05e-05	1.01
DaliLite	12.70	3.27e-05	10.60	0.00192	2021.41
TM-align	22.00	1.85e-12	22.30	5.75e-12	168.32
Sequence	14.50	1.44e-06	16.60	2.1e-08	0.24

Table 1: Summary of method accuracy and running times, for the 24 evaluated protein structural comparison approaches from **Figure 3**. Accuracy of the given approach is shown with respect to its average ranking compared to all considered approaches across all considered real-world PSN sets, and the results are shown based on the area under precision recall curve (AUPR) as well as the area under ROC curve (AUROC). The ranking of each method is expressed as follows. For the given PSN set, we determined which approach results in the highest accuracy (rank 1), the second highest accuracy (rank 2), etc. Then, we averaged the rankings of the given method over all PSN sets. So, the lower the average rank, the better the method. Since our new NormOrderedGraphlet-3-4(K) approach has the best average rank with respect to both AUPR and AUROC (shown in bold), we computed the statistical significance (i.e., *p*-value) of the improvement of NormOrderedGraphlet-3-4(K) over each of the other approaches in terms of their ranks, using paired *t*-test. Running times of the approaches are shown for one of the representative analyzed PSN sets.

The above work was published as follows:
Fazle E. Faisal, Khalique Newaz, Julie L. Chaney, Jun Li, Scott J. Emrich, Patricia L. Clark, and Tijana Milenkovic (2017), **GRAFENE: Graphlet-based alignment-free network approach integrates 3D structural and sequence (residue order) data to improve protein structural comparison**, *Nature Scientific Reports*, 7, Article number: 14890.

2.3 Aim 3: Further algorithmic NA developments

We achieved five accomplishments in this aim.

First, existing NA methods are homogeneous, i.e., they can deal only with networks containing nodes and edges of one type. Due to increasing amounts of heterogeneous network data with nodes or edges of different types, we extended three recent state-of-the-art homogeneous NA methods, WAVE, MAGNA++, and SANA, to allow for heterogeneous NA for the first time. To achieve this, we introduced several algorithmic novelties. Namely, these existing methods compute

homogeneous graphlet-based node similarities and then find high-scoring alignments with respect to these similarities, while simultaneously maximizing the amount of conserved edges. Instead, we extended homogeneous graphlets to their heterogeneous counterparts, which we then used to develop a new measure of heterogeneous node similarity. Also, we extended S^3, a state-of-the-art measure of edge conservation for homogeneous NA, to its heterogeneous counterpart. Then, we found high-scoring alignments with respect to our heterogeneous node similarity and edge conservation measures. In evaluations on synthetic and real-world biological networks, our proposed heterogeneous NA methods led to higher-quality alignments and better robustness to noise in the data than their homogeneous counterparts (**Figure 4**).

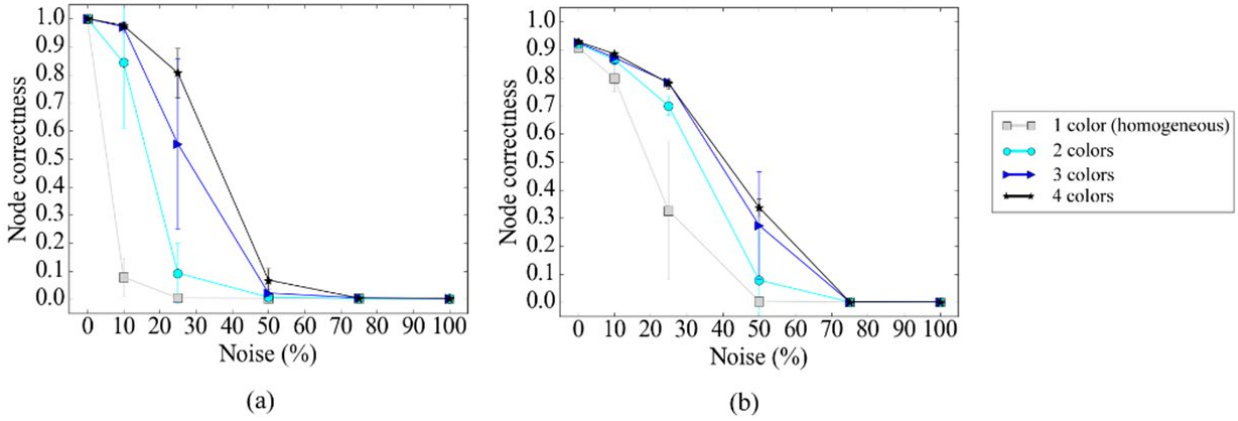


Figure 4: Representative results regarding the effect of the number of considered node colors on alignment quality. 1 color corresponds to traditional homogeneous NA, while 2-4 colors correspond to our new heterogeneous NA; the more node colors are used, the more heterogeneous the given alignment is. Alignment quality is expressed in terms of node correctness, i.e., the percentage of node pairs from the given alignment that are correctly mapped, and it is shown as a function of noise level (percent of rewired edges, varied to be 0%, 10%, 25%, 50%, 75%, and 100%). That is, **(a)** a synthetic network or **(b)** a real molecular (protein-protein interaction (PPI)) network is aligned to its noisy (rewired) counterpart. These results are for using WAVE NA method; results are qualitatively similar when using MAGNA++ or SANA NA methods. For the synthetic network data, nodes are randomly assigned a color out of *k* possible colors, *k* from 1 to 4. That is, for each synthetic network, from the homogeneous (1 color) version, we form heterogeneous versions with 2, 3, and 4 colors. For the real PPI network data, we assign colors to nodes in the 2-colored network based on whether the given protein is aging-related or not; the 3-colored network based on whether the given protein is aging-related only, Alzheimer's disease-related only, or neither; and the 4-colored network based on whether the given protein is aging-related only, Alzheimer's disease-related only, both aging-related and Alzheimer's disease-related, or neither of the two. Clearly, the more colors are used, the higher the alignment quality, which confirms that heterogeneous NA is superior to homogeneous NA.

The above work was published as follows:
Shawn Gu, John Johnson, Fazle E. Faisal, and Tijana Milenkovic (2018), **From homogeneous to heterogeneous network alignment via colored graphlets**, *Nature Scientific Reports,* 8, Article number: 12524.

Second, the problem of NA relies on a subproblem of computing topological similarities between nodes of the aligned networks, which are typically computed via the notion of network embedding. Specifically, network embedding aims to represent each node in a network as a low-dimensional feature vector that summarizes the given node's (extended) network neighborhood. The nodes'

feature vectors can then be used in various downstream machine learning tasks, including computing node similarities. Recently, many embedding methods that automatically learn the features of nodes have emerged, such as node2vec and struc2vec, which have been used in tasks such as node classification, link prediction, and node clustering, mainly in the social network domain. There are also other embedding methods that explicitly look at the connections between nodes, i.e., the nodes' network neighborhoods, such as graphlets. Graphlets have been used in many tasks such as network comparison/alignment, link prediction, and network clustering, mainly in the computational biology domain. Even though the two types of embedding methods (node2vec/struct2vec versus graphlets) have a similar goal – to represent nodes as features vectors, no comparisons have been made between them, possibly because they have originated in the different domains. Therefore, we compared graphlets to node2vec and struc2vec, and we did so in the task of NA. In evaluations on synthetic and real molecular networks, we found that graphlets are both more accurate and faster than node2vec and struc2vec, which will likely result in a wide adoption of graphlets (which were heavily used in this project) in the domain of social networks and thus help increase the visibility of this project.

The above work was published as follows:
Shawn Gu and Tijana Milenkovic (2018), **Graphlets versus node2vec and struc2vec in the task of network alignment**, In Proceedings of the 14th International Workshop on Mining and Learning with Graphs (MLG) at the 24th ACM SIGKDD 2018 Conference on Knowledge Discovery & Data Mining (KDD), London, UK, August 19-23, 2018.

Third, focusing specifically on the application of NA to the computational biology domain, NA in this domain aims to find a node mapping between species' molecular (e.g., protein-protein interaction) networks that uncovers similar network regions, thus allowing for transfer of functional knowledge between the aligned network regions. However, current NA methods do not end up aligning functionally related network regions. A likely reason is that they assume it is topologically similar nodes that are functionally related. However, we showed that this assumption does not hold well – nodes that are topologically similar can be both functionally related and functionally related, and also, nodes that are functionally related can be both topologically similar and topologically dissimilar. So, a paradigm shift is needed with how the NA problem is approached. We proposed such a shift by redefining NA as a data-driven framework, TARA (daTA-dRiven network Alignment), which attempts to learn the relationship between topological relatedness and functional relatedness without assuming that topological relatedness corresponds to topological similarity, like traditional NA methods do. TARA trains a classifier to predict whether two nodes from different networks are functionally related based on their network topological patterns. We found that TARA is able to make accurate predictions. TARA then takes each pair of nodes that are predicted as related to be part of an alignment. Like traditional NA methods, TARA uses this alignment for the across-species transfer of functional knowledge. Clearly, TARA as currently implemented uses topological but not protein sequence information for this task. We find that TARA outperforms existing state-of-the-art NA methods that also use topological information, WAVE and SANA, and even outperforms or complements a state-of-the-art NA method that uses both topological and sequence information, PrimAlign (**Figure 5**). Hence, adding sequence information to TARA in the future is likely to further improve its performance.

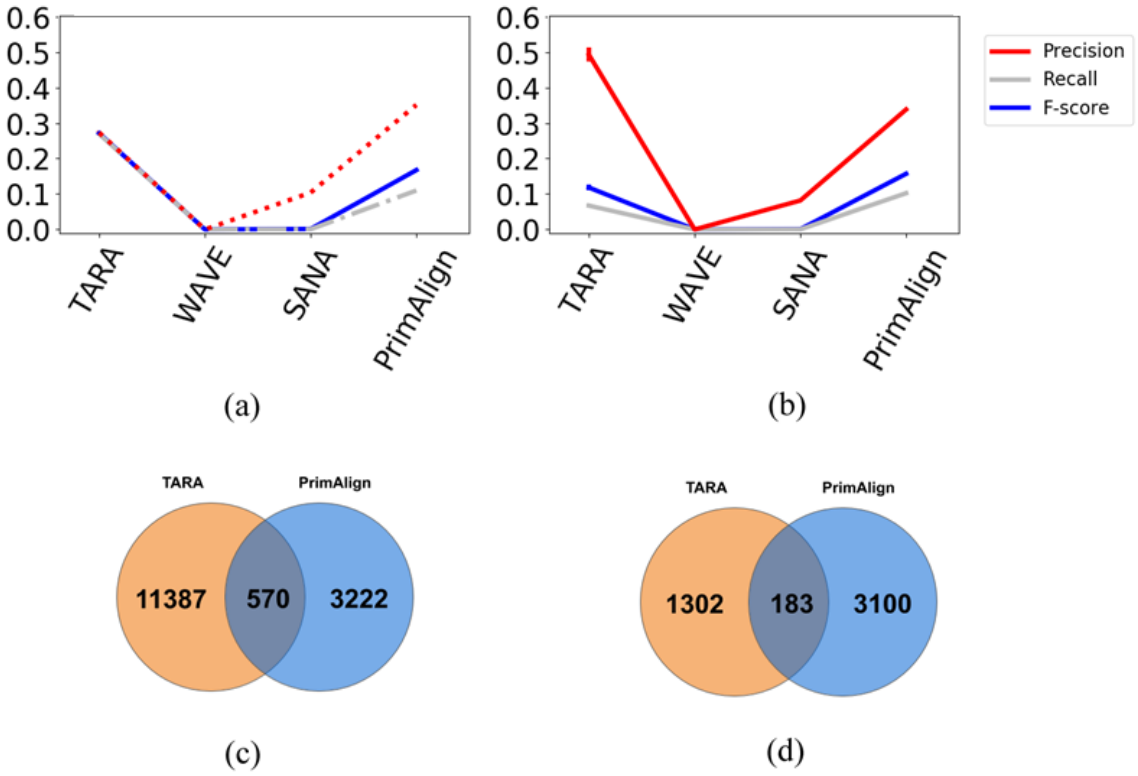


Figure 5: (a)-(b) Alignment accuracy (in terms of precision, recall, and F-score) of our new data-driven TARA approach versus the existing WAVE, SANA, and PrimAlign approaches in the task of protein function prediction from an alignment of yeast and human protein-protein interaction networks. Panels **(a)** and **(b)** differ in which protein functional ground truth dataset we use in our evaluation. Clearly, TARA is always better than WAVE and SANA. Compared to PrimAlign, TARA either has comparable precision but better recall (panel **(a)**) or better precision and comparable recall (panel **(b)**). **(c)-(d)** The numbers of overlapping (i.e., identical) predictions between TARA and PrimAlign (the best of the three existing methods) corresponding to results in panels (a)-(b), respectively. Clearly, the two approaches are highly complementary in which protein-function associations they predict.

The above work has resulted in the following paper:
Shawn Gu and Tijana Milenkovic (2019), **Data-driven network alignment**, under review. Also, arXiv:1902.03277 [q-bio.MN].

Fourth, we have worked on redefining the NA problem as a recommendation system. The latter has been studied in data mining and social network communities for decades. Recommendation systems are widely used to suggest a personalized list of items (e.g., movies, books, or new friends) to individuals based on their preferences to help them find the most relevant items. In other words, for an individual *i*, based on the history of *i*'s behaviors (e.g., rating movies, liking contents on social media, or friendships with other individuals), a recommender system approach calculates a personalized ranking score on a set of new items (i.e., movies, contents, or friends, respectively) and suggests the top-ranked items to *i*. We recognized that the problem of NA – mapping nodes across networks – can be redefined as a recommender system problem. Then, the huge body of research on recommender systems can be used to solve the NA problem. Indeed, this idea is

promising. Namely, in our preliminary evaluation tests, when we aligned four synthetic or molecular networks to their 25% noisy (randomly rewired) versions, meaning that the compared networks are 25% dissimilar, i.e., are 75% similar, and measured node correctness, i.e., the percentage of correctly aligned nodes, our recommender system-based NA approach yielded 57%, 60%, 89%, and 98% accuracy for the four tests, versus existing state-of-the-art NA approaches (WAVE, SANA, and PrimAlign) yielding only up to 16%, 44%, 73%, and 84% accuracy, respectively.

This work on recommendation-based NA, primarily led to date by the project's postdoc, Dr. Fatemeh Vahedian, is in its early stages, and together with TARA, it is expected to act as a set of preliminary results for a follow-up proposal on next-generation network comparison to AFOSR or an alternative government agency.

Fifth, in an applied context, we used a recommender system approach to integrate (rather than compare) into a heterogeneous information network very different data types originating from a rich data set from the University of Notre Dame's NetHealth study that collected individuals' social interaction data via smartphones, health-related behavioral data via wearables (Fitbit), and trait data from surveys. We used the heterogeneous information network to predict likelihood of an individual to be depressed or anxious. Yet, we note that our data integrative framework is generalizable to predicting any other behavior or trait of an individual.

This work has resulted in the following paper:
Shikang Liu, Fatemeh Vahedian, David Hachen, Omar Lizardo, Christian Poellabauer, Aaron Striegel, and Tijana Milenkovic (2019), **Heterogeneous network approach to predict individuals' mental health**, under review. Also, arXiv:1906.04346 [cs.SI].

2.4 Aim 4: NA of dynamic networks

We achieved three accomplishments in this aim.

First, we hypothesized that aligning *dynamic* network representations of evolving systems would produce superior alignments compared to aligning the systems' *static* network representations, as is currently done in the NA field. To test this hypothesis, we introduced the first ever NA method for comparing dynamic networks, DynaMAGNA++. This proof-of-concept dynamic NA method is an extension of a state-of-the-art static NA method, MAGNA++. Even though both MAGNA++ and DynaMAGNA++ optimize edge as well as node conservation across the aligned networks, MAGNA++ conserves static edges and similarity between static node neighborhoods, while DynaMAGNA++ conserves dynamic edges (events) and similarity between evolving node neighborhoods. For this purpose, we introduced the first ever measure of dynamic edge conservation and we relied on our recent existing measure of dynamic node conservation. Importantly, the two dynamic conservation measures can be optimized using any state-of-the-art NA method and not just MAGNA++. We confirmed our hypothesis that dynamic NA is superior to static NA, under fair comparison conditions, on synthetic and real-world networks, in computational biology and social network domains (**Figure 6**). DynaMAGNA++ is parallelized

and it includes a user-friendly graphical interface. All code and data are available at: https://nd.edu/~cone/DynaMAGNA++/ .

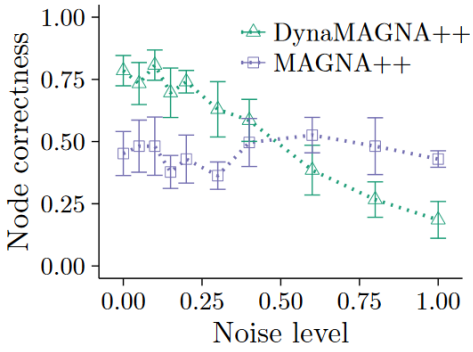


Figure 6: Representative alignment quality (in terms of node correctness) of DynaMAGNA++ and (static) MAGNA++ as a function of noise level when aligning a network to randomized (noisy) versions of the original network. The results are consistent across all analyzed networks. The larger the noise level, the more dissimilar the aligned networks are. Consequently, for a good method, alignment quality should decrease with increase in the noise level, which is the case for DynaMAGNA++ but not for MAGNA++. Note that node correctness of an alignment is the fraction of correctly aligned node pairs (according to the ground truth node mapping) out of all aligned node pairs. Given that our original network and its randomized versions have the same set of nodes, we know which nodes in the original network correspond to which nodes in the given randomized network. That is, we know the ground truth mapping between the aligned networks, and thus, we can measure node correctness. We expect DynaMAGNA++'s alignment quality to be superior to MAGNA++'s alignment quality with respect to node correctness for lower (meaningful) noise levels, if it is indeed true that dynamic NA is superior to static NA. We do not expect this superiority for higher noise levels, since at such noise levels, networks being aligned are highly randomized and thus a good method should produce low-quality alignments. Indeed, our results confirm all of this, i.e., that dynamic NA is superior to static NA.

This work was published as follows:
Vipin Vijayan, Dominic Critchlow, and Tijana Milenkovic (2017), **Alignment of dynamic networks**, *Bioinformatics*, 33(14): i180-i189.

Second, dynaMAGNA++ does not necessarily scale well to larger networks in terms of alignment quality or running time. To address this, more recently, we introduced a new dynamic NA approach, DynaWAVE. Both DynaMAGNA++ and DynaWAVE optimize our new objective functions for finding conserved regions between dynamic (rather than static) networks. The two methods differ mainly in some algorithmic details, specifically in their optimization strategies (how they optimize the new objective functions). On synthetic and real-world (molecular aging-related, ecological animal proximity, and social communication) networks, we found that DynaWAVE complements DynaMAGNA++: DynaMAGNA++ is more accurate than DynaWAVE on smaller networks, while DynaWAVE is more accurate (while also being much faster) than DynaMAGNA++ on larger networks. This justifies the need for both approaches. As more dynamic network data become available, dynamic NA and thus our two methods will continue to gain importance. With this in mind, we provided a friendly user interface and source code for each of the two methods, at https://nd.edu/~cone/DynaMAGNA++/ and https://nd.edu/~cone/DynaWAVE/ , respectively.

The DynaWAVE work was published as follows:
Vipin Vijayan and Tijana Milenkovic (2017), **Aligning dynamic networks with DynaWAVE**, *Bioinformatics*, 34(10): 1795–1798.

Third, recall that NA methods optimize node conservation and edge conservation. Different node conservation measures exist. Dynamic graphlet degree vectors are a state-of-the-art dynamic node conservation measure, used within both DynaMAGNA++ DynaWAVE. Most recently, we used graphlet-orbit transitions (GoTs), a different graphlet-based measure of temporal node similarity, as a new dynamic node conservation measure within DynaWAVE, resulting in a new dynamic NA approach called GoT-WAVE. On synthetic networks, GoT-WAVE improved DynaWAVE's accuracy by 30% and speed by 64%. On real networks, when optimizing only dynamic node conservation, the two methods are complementary – each is better than the other one for some of the analyzed networks. However, only GoT-WAVE supports directed edges. Hence, GoT-WAVE is a promising new dynamic NA algorithm. We provided a user-friendly user interface and source code for GoT-WAVE at http://www.dcc.fc.up.pt/got-wave/ (note that this is a collaborative project and it is the collaborators who are hosting the GoT-WAVE web site).

This work was published as follows:
David Aparicio, Pedro Ribeiro, Tijana Milenkovic, and Fernando Silva (2019), **Temporal network alignment via GoT-WAVE**, *Bioinformatics*, doi: 10.1093/bioinformatics/btz119.

AFOSR Deliverables Submission Survey

Response ID:11767 Data

1.

Report Type

Final Report

Primary Contact Email

Contact email if there is a problem with the report.

tmilenko@nd.edu

Primary Contact Phone Number

Contact phone number if there is a problem with the report

5746318975

Organization / Institution name

University of Notre Dame

Grant/Contract Title

The full title of the funded effort.

(YIP) Efficient Comparison of Multiple Complex Networks

Grant/Contract Number

AFOSR assigned control number. It must begin with "FA9550" or "F49620" or "FA2386".

FA9550-16-1-0147

Principal Investigator Name

The full name of the principal investigator on the grant or contract.

Tijana Milenkovic

Program Officer

The AFOSR Program Officer currently assigned to the award

Tristan Nguyen

Reporting Period Start Date

07/01/2016

Reporting Period End Date

06/30/2019

Abstract

Network alignment (NA), one of the most popular network science/mining tasks, aims to compare networks corresponding to different systems in order to identify network regions of the systems' (dis)similarities, thus allowing for learning something about a poorly understood system from a well understood system based on their aligned network regions. As such, NA has applications in a variety of domains, including computational biology, chemoinformatics, neuroscience, computational linguistics, artificial intelligence, computer vision, and web mining. Since complexity theory dictates that the problem of NA is computationally hard, this project introduced novel computationally efficient yet accurate heuristic NA approaches, such as those for alignment of
multiple networks (as opposed to traditional pairwise NA), dynamic networks (as opposed to traditional static NA), or

heterogeneous networks (as opposed to traditional homogeneous NA). The project resulted in 10 published or submitted papers and 16 conference presentations of the project results. It supported eight researchers (the principal investigator, a postdoctoral researcher, four Ph.D. students, and two undergraduate students).

Distribution Statement

This is block 12 on the SF298 form.

Distribution A - Approved for Public Release

Explanation for Distribution Statement

If this is not approved for public release, please provide a short explanation. E.g., contains proprietary information.

SF298 Form

Please attach your SF298 form. A blank SF298 can be found here. Please do not password protect or secure the PDF The maximum file size for an SF298 is 50MB.

SF_298.pdf

Upload the Report Document. File must be a PDF. Please do not password protect or secure the PDF . The maximum file size for the Report Document is 50MB.

AFOSR_YIP_final_report.pdf

Upload a Report Document, if any. The maximum file size for the Report Document is 50MB.

Archival Publications (published) during reporting period:

Vipin Vijayan and Tijana Milenkovic (2017), Multiple network alignment via multiMAGNA++, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15(5): 1669-1682.

Fazle E. Faisal, Khalique Newaz, Julie L. Chaney, Jun Li, Scott J. Emrich, Patricia L. Clark, and Tijana Milenkovic (2017), GRAFENE: Graphlet-based alignment-free network approach integrates 3D structural and sequence (residue order) data to improve protein structural comparison, Nature Scientific Reports, 7, Article number: 14890.

Vipin Vijayan, Dominic Critchlow, and Tijana Milenkovic (2017), Alignment of dynamic networks, Bioinformatics, 33(14): i180-i189.

Vipin Vijayan and Tijana Milenkovic (2018), Aligning dynamic networks with DynaWAVE, Bioinformatics, 34(10): 1795–1798.

Shawn Gu, John Johnson, Fazle E. Faisal, and Tijana Milenkovic (2018), From homogeneous to heterogeneous network alignment via colored graphlets, Nature Scientific Reports, 8, Article number: 12524.

Shawn Gu and Tijana Milenkovic (2018), Graphlets versus node2vec and struc2vec in the task of network alignment, In Proceedings of the 14th International Workshop on Mining and Learning with Graphs (MLG) at the 24th ACM SIGKDD 2018 Conference on Knowledge Discovery & Data Mining (KDD), London, UK, August 19-23, 2018.

David Aparicio, Pedro Ribeiro, Tijana Milenkovic, and Fernando Silva (2019), Temporal network alignment via GoT-WAVE, Bioinformatics, doi: 10.1093/bioinformatics/btz119.

Vipin Vijayan, Eric Krebs, and Tijana Milenkovic (2019), Pairwise versus multiple network alignment, under review. Also, arXiv:1709.04564 [q-bio.MN].

Shawn Gu and Tijana Milenkovic (2019), Data-driven network alignment, under review. Also, arXiv:1902.03277 [q-bio.MN].

Shikang Liu, Fatemeh Vahedian, David Hachen, Omar Lizardo, Christian Poellabauer, Aaron Striegel, and Tijana Milenkovic (2019), Heterogeneous network approach to predict individuals' mental health, under review. Also, arXiv:1906.04346 [cs.SI].

New discoveries, inventions, or patent disclosures:

Do you have any discoveries, inventions, or patent disclosures to report for this period?

No

Please describe and include any notable dates

Do you plan to pursue a claim for personal or organizational intellectual property?

Changes in research objectives (if any):

N/A

Change in AFOSR Program Officer, if any:

My originally assigned Program Officer was Dr. James Lawton.

In project year 2 (2017/2018), Dr. Tristan Nguyen has taken over the role of the Program Officer for this award.

Extensions granted or milestones slipped, if any:

N/A

AFOSR LRIR Number

LRIR Title

Reporting Period

Laboratory Task Manager

Program Officer

Research Objectives

Technical Summary

Funding Summary by Cost Category (by FY, $K)

	Starting FY	FY+1	FY+2
Salary			
Equipment/Facilities			
Supplies			
Total			

Report Document

Report Document - Text Analysis

Report Document - Text Analysis

Appendix Documents

2. Thank You

E-mail user

Sep 03, 2019 18:44:20 Success: Email Sent to: tmilenko@nd.edu