Detecting Anomalies in Spatiotemporal Data using Genetic Algorithms with Fuzzy Community Membership

Garnett Wilson*, Simon Harding*, Orland Hoeber*, Rodolphe Devillers† and Wolfgang Banzhaf*
*Department of Computer Science
Memorial University of Newfoundland, St. John's, Canada A1B 3X5
Email: gwilson, simonh, hoeber, banzhaf@mun.ca
†Department of Geography
Memorial University of Newfoundland, St. John's, Canada A1B 3X5
Email: rdeville@mun.ca

***Abstract*—A genetic algorithm is combined with two variants of the modularity (Q) network analysis metric to examine a substantial amount fisheries catch data. The data set produces one of the largest networks evaluated to date by genetic algorithms applied to network community analysis. Rather than use GA to decide community structure that simply maximizes modularity of a network, as is typical, we use two fuzzy community membership functions applied to natural temporal divisions in the network so the GA is used to find interesting areas of the search space through maximization of modularity. The work examines the performance of the genetic algorithm against simulated annealing using both types of fuzzy community membership functions. The algorithms are used in an existing visualization software prototype, where the solutions are evaluated by a fisheries expert.**

***Keywords*-spatiotemporal visualization; fisheries; genetic algorithm; Q modularity; fuzzy community membership**

I. Introduction

A number of works have combined genetic algorithm (GA) network search with the modularity (or Q) metric that rewards higher connection density within communities and sparseness between communities. Typically the GA itself is used to create a community structure that maximizes Q. If the network is very large, the evaluation of Q for each GA solution that proposes community division of the network rapidly becomes prohibitive and thus large networks are seldom used. Rather than use GA with Q to decide community structure of a network, we use GA guided by Q-based fitness based on inherent community structure to search for interesting relationships in the data set. The network examined is based on fisheries catch data compiled by the Canadian Department of Fisheries and Oceans (DFO) for the Newfoundland and Labrador region, and is the largest (to the authors' knowledge) network yet evaluated using GA for network community analysis. To leverage inherent community structure, we allow overlapping of communities using two variants of a fuzzy community membership function in the traditional Q modularity metric.

Section 2 of this paper discusses background including the mathematics behind determining modularity (community structure, or Q) of a network using two variants of fuzzy community membership, and it also explores previous applications of GAs to network analysis. Section 3 describes the network-based interpretation of the spatiotemporal catch data and how solutions are visualized using existing software called "GTdiff." Section 4 describes the GA algorithm used for fuzzy community analysis. Section 5 examines the performance results of the GA compared to simulated annealing (SA) using both types of fuzzy community membership, with conclusions following in Section 6.

II. Background and Related Work

A. Determination of Network Community Structure

This paper uses metrics for graph/social network analysis (SNA) to automatically identify events associated with the collapse of the Newfoundland cod fishery in the early 1990s. As will be described in greater detail in Section 3, the GA evaluates networks consisting of nodes and edges based on their community structure. The nodes are uniquely identified by both spatial position and their time span; weights are assigned to edges based on differences between the node information. A popular metric for evaluation of community structure in a network, more sophisticated than other basic graph theoretic measures such as density or degree, is the *modularity* (or Q) metric. Practically speaking, large Q values indicate a network with dense internal connections between the nodes within communities but sparse connections between separate communities. Newman adapted the modularity metric for weighted networks [1]; the importance of weight on community structure in networks is substantiated by Fan et al. [2]. The Q^w metric we use, slightly adapted from [1], is defined as

$$Q_w = \frac{1}{2m} \sum_{ij} \left(A_{ij} - \frac{k_i k_j}{2m} \right) \delta_f(c_i, c_j) \qquad (1)$$

where A_{ij} is the weight of the connection from i to j, k_i of a node i in a weighted network is the sum of the weights of the edges attached to it ($k_i = \sum_j A_{ij}$), and $m = \frac{1}{2} \sum_{ij} A_{ij}$ is

the number of edges in the network. Q_w can take an absolute value between 0 and 1, with values of over 0.3 indicating good community division of the network [1].

The function is a fuzzy version of the more traditional community membership function $\delta_f(c_i, c_j)$, where c_i is the community to which the node i is assigned. In the traditional (non-fuzzy) community membership function δ, communities do not overlap: that is, any particular node cannot be a member of more than one community. In this work, it is not appropriate to consider time spans as individual, non-overlapping communities because if two nodes possessed the same time span no difference in data would be represented. Edges in the network using two nodes with the same time span are thus prohibited since they do not reflect any difference in catch due to the entire geographical area being used in the final visualization (discussed in Section 3). The practical implication of this restriction, since nodes are identified by both location and time span, is that loops (reflexive ties) are prohibited.

We calculate fuzzy community membership as the degree of overlap between the time spans within each of two nodes:

$$\delta_f(c_i, c_j) = \frac{|Y_i \cap Y_j|}{|Y_i \cup Y_j|} \tag{2}$$

where Y_i is the enumeration of years, inclusive, for the time span of node i. Similarly, Y_j is the enumeration of years for the time span corresponding to node j. The function δ_f returns a decimal value between 0 and 1 and thus maintains the natural range of values for Q_w. In practice, the δ_f function tends to cause Q_w to favor networks with edges using overlapping time spans so that the user is exposed to larger differences within the overlapping time frames. For instance, the user may see that two years out of a 10 year span involve abnormally low catches on average. Another variant of Equation 1 is used where δ_f is replaced by $1-\delta_f$ so that non-overlapping time spans are favored. With this variation, the user is more likely to be presented with large differences that occurred between independent time spans. For instance, the user may be presented with a 2 year span where (on average) catches went up compared to a 5 year span a few years earlier. In this work, we compare a genetic algorithm and simulated annealing using the Q_w variants just described.

B. Use of GAs for Community Detection

Genetic algorithms are a popular means of optimizing Q_w, especially if a large network that cannot be handled by more conventional (and exhaustive) search techniques is being considered. Tasgin et al. [3] applied a genetic algorithm to the assignment of communities to network nodes where the fitness function of the GA was the optimization of the Q metric, with a GA individual consisting of a mapping of each individual node to a community. Tasgin et al. note that the use of a GA precludes the need for *a priori* specification of communities, and it works well and is scalable for large networks. Gog et al. [4] use a GA algorithm modified so that individuals are collaborative in the sense that they are aware of the global optimum solution obtained at any time and their best ancestor. The GA individuals used dictate the mapping of each node in a network to a particular community. Liu et al. [5] use a GA to repeatedly subdivide a network into communities and correctly assign individuals to communities. Shi et al. [6] introduce a Q metric-based GA that uses adjacency information in the genotype to reduce the search space compared to other community division algorithms with subsequent modularity evaluation, with the added benefit that the number of communities need not be known or preselected. Nicosia et al. [7] present the use of a GA with overlapping community structure incorporated in the Q metric, and demonstrate its benefit on a few small to moderately sized networks.

The Q metric is not the only choice for fitness function for a GA search. Rather than explicitly optimizing the Q metric as a fitness function, Pizzuti [8] attempts to identify densely connected groups of nodes separated by sparse connections using a fitness function called community score to identify the community structure of a network. Firat et al. [9] do not use the Q metric for fitness either; instead they use a random walk distance measure between cluster centers as nodes with the number of clusters decided *a priori*.

Most of the works above attempt to optimize Q to establish optimal divisions of often-used networks in the social network literature. For smaller networks with known community structure, the authors check the accuracy of their algorithms. For larger networks without established underlying community structure, the authors often attempt to establish higher modularity measures for similar computational effort when competing with other algorithms. The goal of this work is different in at least two major respects. Firstly, we use a very large real world data set that has never been examined with GAs or subjected to social network analysis (including community-based modularity). Secondly, we do not attempt to impose community structure on the data set and then maximize modularity; instead, we attempt to use modularity as a fitness function to find pertinent relationships using the natural temporal divisions considered as pre-established communities.

III. Analysis of Spatiotemporal Data

A. Network Interpretation of Spatiotemporal Catch Data

The network used is based on a spatiotemporal data set of annual bottom trawl survey catch data for the Atlantic cod (*Gadus morhua*) conducted by the Canadian Department of Fisheries and Oceans (DFO) for the Newfoundland and Labrador, Canada region across 1,000,000 km^2 and over a temporal range of 1980–2005. The data set produces a very large network to be evaluated. The algorithm for the search space includes a node for every combination of spatial point

in an N x N grid and two year time span. The number of unique two year time spans considered for 25 years (1980 to 2005, inclusive) is $\binom{26}{2}$, or 325 possibilities. We also consider the span of one year (e.g. 1996–1996) as a time span, so the the number of possible time spans is 325 + 26 = 351 in total. That is, the years of two time spans can overlap, but the years of the single time span for each node must be unique (cannot refer to one individual year). Each node is identified by both a location x, y and time span. If we restrict the search space to a 30 x 30 grid (shown to be an appropriate resolution for viewing changes in preliminary experiments with an expert) for each possible time span, then there are 30^2 x 351 = 315, 900 nodes to consider.

In practical terms, each one of the edges in the network corresponds to a difference in catch data between two areas over two time spans that could be of interest to an expert user. We wish to consider the difference between nodes (average catch data over a particular area during a time span) as absolute differences. Thus, the network to be considered consists of an undirected, weighted graph. The number of all unique edges existing in this search space is the number of possible pairings of nodes, with no time span compared to itself for a difference of 0, giving $n(n-1)/t$ for n nodes and t time spans, or approximately 2.8 x 10^8 possible edges. A node and edge with corresponding spatiotemporal representation are shown in Fig. 1.

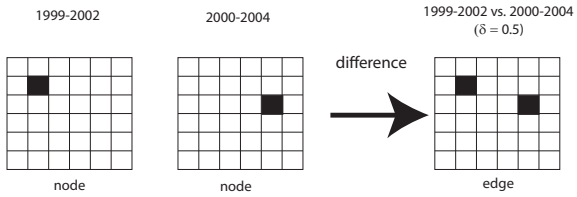


Figure 1. Relationship between network structure and spatiotemporal visualization.

B. Spatiotemporal Visualization System

The goal of the search algorithms (GA and SA) are to automatically identify the conditions under which there are significant changes in the spatiotemporal data. This can be used to automatically configure the spatial and temporal settings of a visualization system designed for the exploration of such data. The approach can direct the user to interesting features that emerge from the data, which can then be explored further within a visualization system. An example of such a visualization system is GTdiff [10], which was designed to allow users to see patterns in how data has changed over time. In the current system, users explore the data manually and may start by filtering the data temporally. Within the specified temporal range, the data can then be grouped into a user-specified number of temporal bins. For instance, users may find it useful to examine data only over a particular five year period grouped into five 1 year bins, or over 16 years of data grouped into four 4 year bins.

Simultaneously, the data within any given temporal bin is also placed into spatial bins. The spatial binning of the data is necessary in order to address minor spatial variations in the data. Particular spatial bins are compared in GTdiff using a grid of N x N units, with each unit of the grid displaying an average of the data point samples within the particular unit of the grid. In addition to showing the set of spatial bins for an individual temporal bin, GTdiff provides a set of difference graphs that visually highlight the changes in the associated temporal bins using shades of green for positive changes and shades of red for negative changes.

The networks which the GA individuals represent naturally map to the display features of GTdiff. In particular, the network nodes each correspond to an x, y point on an N x N grid for a particular span of years (i.e., a temporal bin). Iterating through each valued node in a network which the GA evolved will produce a list of time spans (temporal bins) to be displayed to the user. The edges in the network evolved by the GA thus correspond to differences between the time spans, which are visualized in GTdiff as the difference graphs seen in Figures 6 to 7 (discussed in Section 5).

IV. GA Fuzzy Community Algorithm

Each GA individual genotype consists of 20 chromosomes, where a chromosome is an ordered set of 8 integers identifying an edge in a network. The first 4 integers identify a node on one end of the edge, while the last 4 integers represent the node at the other end of the edge. For each set of 4 integers corresponding to a node, the first two integers identify the x and y co-ordinates in the N x N grid and the last two integers identify the start and end years of a time span in the data set. The integer corresponding to the end year of each time span is naturally restricted so that it is not greater than the initial year, and (as discussed previously) the time spans cannot be the same for a difference of 0. The absolute difference between the average catch over all years for the location in each node is the weight of the edge. The chromosome of the GA individual representing an edge is:

$$\overbrace{\overbrace{\underbrace{x_1, y_1}_{\text{location1}}\ ,\ \underbrace{t_1, t_2}_{\text{timespan1}}}^{\text{node1}}\ \overbrace{\underbrace{x_2, y_2}_{\text{location2}}\ ,\ \underbrace{t_3, t_4}_{\text{timespan2}}}^{\text{node2}}}^{\text{edge}} \tag{3}$$

where $t_2 \geq t_1$, $t_4 \geq t_3$ and $t_1, t_2 \neq t_3, t_4$. The GA individuals, being a list of edges, represent potential networks of interest and are evaluated in a steady state tournament of 100,000 rounds using a population of 10. The small population size is used so that the process of evolutionary search will guide the construction of interesting networks rather than rely on the possibility that randomly generated material exists in a larger initial population to be discovered

through extensive search. At each round in the tournament, four individuals are selected for evaluation. The top two individuals are kept untouched (become "parents"), while the losing two individuals are replaced by copies of the genotype of the two winners (become the "children"). The copies of the two winners are then subjected to the genetic operations of mutation and crossover.

The mutation operator is always invoked on the copied individuals (children), but each chromosome has a 50% (rate of 0.5) chance of being mutated so some nodes and edges copied from the parent are retained and others are replaced with newly generated edges in order to explore the search space. The crossover operator exchanges two equally sized portions of the two children genotypes, where the size of the portions exchanged is less than the maximum number of chromosomes (20 in this work). The crossover of segments occurs 50% of the time (the crossover rate is 0.5). The fitness function used to evaluate the GA individuals is a fuzzy Q_w that either incorporates preference for time span overlap (δ_f) or preference for no overlap ($1 - \delta_f$). We also compare the GA to simulated annealing (SA) using all these fitness metrics, where SA is often considered the most optimum algorithm for accurate community detection in large search spaces [5].

V. Results

A. Quantitative Results

The GA and SA algorithms were both run for an equal number of evaluations: since the GA processed four individuals per each of 10 000 tournament rounds, the SA was permitted to run for 40 000 cycles. The SA algorithm keeps track of the best state found so far, as well as a current state in the search. Each cycle of the SA replaces the current state with a new candidate state with probability $e^{\Delta E/T}$ where ΔE is a change in the value of Q and T is the current temperature of the system. The SA replaces the best state found so far if the new candidate state has a higher Q value. The temperature reduction schedule corresponds directly to the number of cycles completed. The quantitative results at the end of GA and SA execution are shown in boxplots in Figures 2 through 5 for 50 trials (of 40 000 evaluations each). Bottom, middle, and top of boxes indicate lower quartile, median, and upper quartile values, respectively. If notches of boxes do not overlap, medians of the two sets of data differ at the 0.95 confidence interval. The symbol '+' denotes points from 1.5 to 3 times the interquartile range, and 'o' denotes points outside 3 times the interquartile range.

Fig. 2 shows the best Q-based fitness achieved using fuzzy community membership functions with overlapping favored using δ_f (OF) and non-overlapping favored using $1-\delta_f$ (NOF). The number of communities in the networks corresponding to these best Q fitness values are shown in Fig. 3. Maximum difference present in each of the best networks is shown in Fig. 4, with the time required to evaluate all networks in each round/iteration for the GA/SA (respectively) shown in Fig. 5. Fig. 2 demonstrates that the GA was able to produce solutions with higher Q fitness (for each of overlapping and non-overlapping favored) than simulated annealing (SA) by a considerable margin. The results for both the GA and SA implementations show that Q fitness using non-overlapping favored community membership (Q, NOF) yields considerably higher Q results than overlapping favored community membership (Q, OF). Fig. 3 shows that both the GA and SA create networks

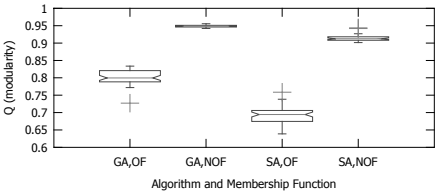


Figure 2. Best modularity (Q) networks located by GA and SA using overlapping favored (OF) and no overlapping favored (NOF) community membership over 50 trials.

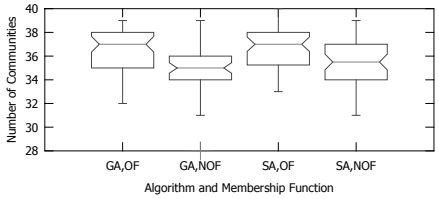


Figure 3. Number of communities within the best modularity (Q) networks located by GA and SA using overlapping favored (OF) and no overlapping favored (NOF) community membership over 50 trials.

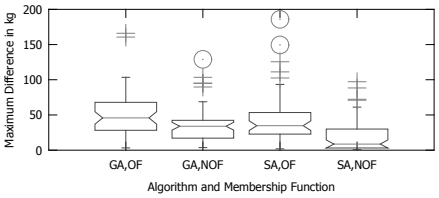


Figure 4. Maximum difference within the best modularity (Q) networks located by GA and SA using overlapping favored (OF) and no overlapping favored (NOF) community membership over 50 trials.

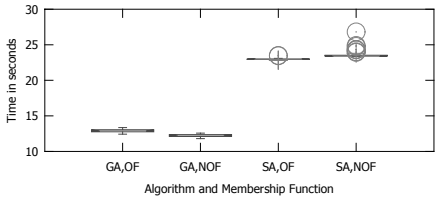


Figure 5. Time per trial (40 000 evaluations) for GA and SA using overlapping favored (OF) and no overlapping favored (NOF) community membership over 50 trials.

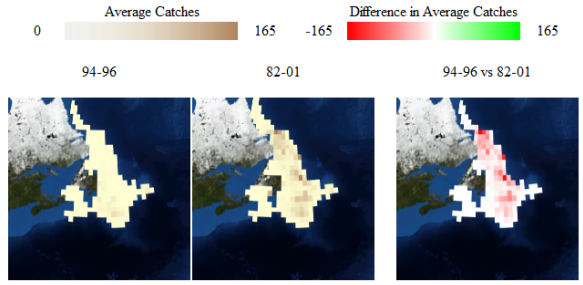
involving fewer communities using the overlapping favored function (with a total of 40 communities being possible), which means they succeed in creating a tighter community structure than favoring non-overlapped communities.

In terms of the maximum difference (Fig. 4) that is located in the networks of high modularity, it is interesting that there is no statistical difference between the GA and SA (in contrast to other metrics examined) for all scenarios except for the poor performance of SA, NOF. However, in terms of overall spread of the data, the GA incorporating an overlap-favoring community membership function provides larger differences than SA, and OF outperforms NOF for both GA and SA. The time in seconds per trial is provided in Fig. 5 using 64-bit Windows 7 Ultimate with a 2.8 GHz Intel Core 2 Duo with 4GB RAM, and each trial consisted of 40 000 evaluations of individual networks for both GA and SA. Comparing GA to SA performance time, the GA is able to more quickly evaluate networks (is typically approximately 10 seconds faster) than SA per trial, but both techniques thus provide reasonable run times for a practical user-centered interactive system like GTdiff.

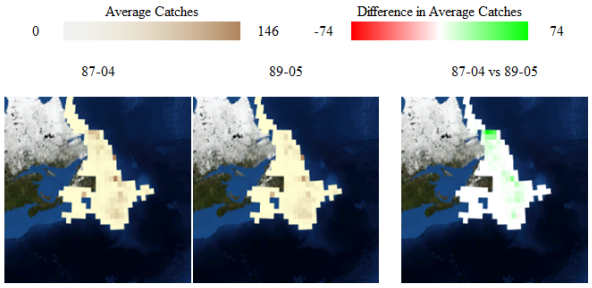
B. Anomaly Detection

Once the networks containing anomalies (where "anomalies" are considered interesting features of the data) are located using the search algorithms, they are visualized using GTdiff as triples (two temporal bins and one corresponding difference graph). Figures 6 to 7 show the most significant differences produced in the final graphs for the overlap favored and no overlap favored (respectively) community membership functions for GA and SA as chosen by our project's fisheries expert (fourth author). The first two grids in each figure correspond to temporal bins in GTdiff, with the average catch in kg displayed in each spatial grid element. The first two temporal bins are always ordered sequentially by the last year of the time spans, or if the last year is the same, by the first year. The colour scale runs from light yellow (lowest average catch) to brown (largest average catch). The difference graph is displayed in the third grid, with the difference in average catch across the two time spans displayed as a positive (green) or negative (red) change. White represents no change in catch, the degree of saturation of green is used to represent positive differences, and the degree of saturation of red is used to represent negative differences.

Biologists reported that cod population levels dropped suddenly in the early 1990s, which prompted a moratorium on cod fisheries from 1992 to 1993. According to our fisheries expert, the salient difference for the combination of GA and overlap-favored community membership (Fig. 6a) corresponds to 1994–1996 compared to 1982–2001. These two time spans overlap; indeed, 1994–1996 is contained completely within 1982–2001. This example clearly shows that the years 1994–1996 following the biological phenomenon

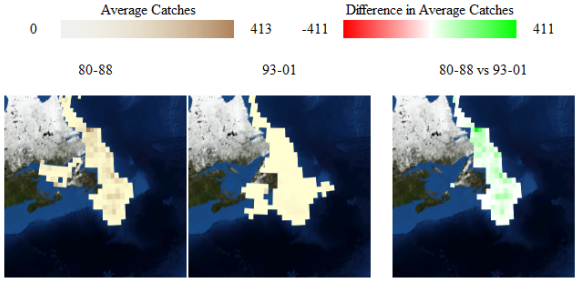


(a) Difference graph located by GA

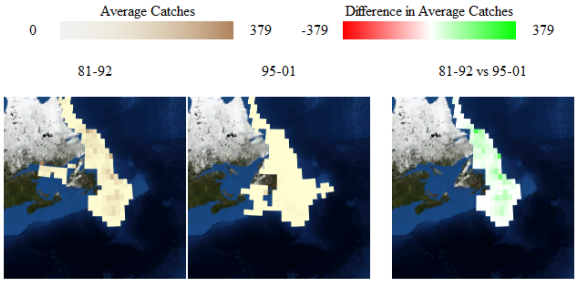


(b) Difference graph located by SA

Figure 6. Difference graph selected by expert from the highest Q, overlap favored network produced by GA (top) and SA (bottom).



(a) Difference graph located by GA



(b) Difference graph located by SA

Figure 7. Difference graph selected by expert from the highest Q, no overlap favored network produced by GA (top) and SA (bottom).

of the collapse (1989–1991) and political moratorium on the cod fishery in NL (1992–1993) involved substantially reduced catch samples (shown in red throughout the grid) compared to the larger time period stretched across most years of the data set (1982–2001 out of 1980–2005). In terms of differences with the graphs produced by simulated annealing, the fisheries expert identified the time spans 1987–2004 to 1995–2001 as being of interest in the SA favoring overlap in communities (Fig. 6b). Rather than picking out clear differences associated with the cod population collapse and moratorium, the difference corresponds to a rapid drop in cod catch in the northern part of the grid (darker green) known to have occurred in this region following 1987–1989 (making the catch samples for the time span including those years appear larger due to the subsequent drop in catch).

The fisheries expert was most impressed with the discovery of the salient difference present in the solution provided by the GA with non-overlapping spans favored in community membership (Fig. 7a). The difference presented was 1980–1988 compared to 1993–2001, where 1988 is considered the last year before the biological collapse of the cod leading to the moratorium and 1993 is the first year following the moratorium. The difference graph shows that stocks were much higher (green throughout grid) previous to the collapse and after the moratorium. SA favoring non-overlap (Fig. 7b) in communities produced the difference 1981–1992 to 1995–2001 that showed higher catch from the start of the survey years (1981) for a number of good years up to (and including) the biological collapse compared to post-moratorium years 1995–2001. However, according to the expert, the separation of cod catch difference between the collapse/moratorium and the preceding better years was not as evident in this example as it was in the differences selected by the GA (Fig. 6a and Fig. 7a).

VI. Conclusion

This work examined the application of a genetic algorithm for examining a very large network space, and applied it to data analysis in a real world system to be used by fisheries experts to examine significant changes over time and location. Experiments showed the GA significantly outperformed the SA in terms of locating the highest modularity-based fitness networks with respect to two fuzzy community membership functions. The overall maximum difference in catch located in the best networks was found in networks where the fuzzy overlap-favoring membership function was used for both GA and SA. Both search algorithms (GA and SA) provided acceptable search times for software that will produce information for users in actual practice, but the GA outperformed SA. Future work will examine alternative fitness functions for the GA evaluation of networks. Also, we plan to compare an arbitrarily assigned community membership function to the use of pre-existing temporal community divisions (as used in this work) using a co-evolutionary system that co-evolves candidate networks along with mappings of nodes to communities.

Acknowledgment

The authors wish to thank Fisheries and Oceans Canada (DFO) for making available the data used in the case study. This work was supported by a Strategic Projects Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC) held by O.H., R.D., and W.B.

References

[1] M. E. J. Newman, "Analysis of weighted networks," *Phys. Rev. E*, vol. 70, no. 5, p. 056131, Nov 2004.

[2] Y. Fan, M. Li, P. Zhang, J. Wu, and Z. Di, "The effect of weight on community structure of networks," *Physica A: Statistical Mechanics and its Applications*, vol. 378, no. 2, pp. 583 – 590, 2007.

[3] M. Tasgin and H. Bingol. (2006, Apr) Community detection in complex networks using genetic algorithm. [Online]. Available: http://arxiv.org/abs/cond-mat/0604419

[4] A. Gog, D. Dumitrescu, and B. Hirsbrunner, "Community detection in complex networks using collaborative evolutionary algorithms," in *ECAL'07: Proceedings of the 9th European Conference on Advances in Artificial Life*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 886–894.

[5] X. Liu, D. Li, S. Wang, and Z. Tao, "Effective algorithm for detecting community structure in complex networks based on GA and clustering," in *ICCS '07: Proceedings of the 7th International Conference on Computational Science, Part II*. Berlin, Heidelberg: Springer-Verlag, 2007, pp. 657–664.

[6] C. Shi, Y. Wang, B. Wu, and C. Zhong, "A new genetic algorithm for community detection," in *Complex Sciences*, ser. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 5. Springer Berlin Heidelberg, 2009, pp. 1298–1309.

[7] V. Nicosia, G. Mangioni, V. Carchiolo, and M. Malgeri, "Extending the definition of modularity to directed graphs with overlapping communities," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2009, no. 03, 2009.

[8] C. Pizzuti, "GA-Net: A genetic algorithm for community detection in social networks," in *Proceedings of the 10th International Conference on Parallel Problem Solving from Nature*. Berlin, Heidelberg: Springer-Verlag, 2008, pp. 1081–1090.

[9] A. Firat, S. Chatterjee, and M. Yilmaz, "Genetic clustering of social networks using random walks," *Comput. Stat. Data Anal.*, vol. 51, no. 12, pp. 6285–6294, 2007.

[10] O. Hoeber, G. Wilson, S. Harding, R. Enguehard, and R. Devillers, "Exploring geo-temporal differences using GTdiff," submitted to *GIS 2010: 18th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems* under review. New York, USA: ACM Press, 2010.