The Meaning of Dissimilar: An Evaluation of Various Similarity Quantification Approaches Used to Evaluate Community Detection Solutions

1st Obaida Hanteer
Data Science & Society Lab
IT University Of Copenhagen
Copenhagen, Denmark
obha@itu.dk

2nd Luca Rossi
Data Science & Society Lab
IT University Of Copenhagen
Copenhagen, Denmark
lucr@itu.dk

***Abstract*—Evaluating a community detection method involves measuring the extent to which the resulted solution, i.e clustering, is similar to an optimal solution, a ground truth. Different normalized similarity indices have been proposed in the literature to quantify the level of similarity among two clusterings where 1 refers to complete agreement and 0 refers to complete disagreement between them. While interpreting the similarity score of 1 seems to be intuitive, it does not seem to be so when the similarity score is otherwise (including 0) suggesting a level of disagreement between the compared clusterings. That is because the there is no agreed upon definition of what is considered dissimilar when it comes to comparing two clusterings. In this paper, we address this issue by first providing a taxonomy of the similarity indices commonly used for evaluating community detection solutions. We then elaborate on the meaning of clusterings dissimilarity and the types of possible dissimilarities that can exist among two clusterings in the context of community detection. We perform an extensive evaluation to study the behaviour of the different similarity indices as a function of the dissimilarity type with both disjoint and non-disjoint clusterings. We finally provide practitioners with some insights on which similarity indices to use for the task at hand and how to interpret their similarity scores.**

I. Introduction

A core task in network analysis is to identify communities; that is, to provide a clustering of the network nodes into groups of nodes, also known as communities or clusters, based on the explicit ties among these nodes and/or some common attributes. Since members of a community tend to generally share common properties, revealing the community structure can provide a better understanding of the overall functioning of the network. This had many applications in social media group detection (1), second-order flow analysis in human mobility (2), and topic detection in information networks (3), just to cite a few.

Lots of efforts have been devoted in the last decade on finding graph-like models for the different types of interactive systems ranging from the very simple (one type of relationships) to the very complex (multiple types of relationships and actors) (4). Accordingly, new community detection methods had to be developed to cope with the different levels of complexity in the different models and this resulted on a large number community detection, clustering, algorithms which might differ from one another on how they define a community structure, but no matter how differ in their complexity they all agree on the structure of the output they produce, that is a clustering $\mathcal{C} == \{C_1, C_2, \ldots, C_k\}$ where C_1, C_2, …, C_k are communities, clusters, over the node-set (the vertex-set that constitutes the input graphs). For the rest of this papers we might interchangeably use the term *clustering* and *community detection solution* to refer to the same thing.

The growing number of community detection methods resulted on an urgent need for tools to evaluate the predicted clusterings of these methods such that we can either compare them with a reference clustering, a ground truth, or we compare the different community detection solutions among each other. For this goal, multiple normalized similarity indices have been borrowed from other proposed or borrowed disciplines to quantify the level of agreement between two clusterings such that the score is 1 when the two clusterings are identical and 0 when the two clusterings are dissimilar. While the meaning of similar seems to be quite intuitive when the similarity score is 1, it is quite unclear what is perceived as less similar or completely dissimilar (when the score is 0) according to these similarity indices. In this paper, we address this issue by first providing a taxonomy of the similarity indices commonly used for evaluating community detection solutions. We elaborate on the meaning of clusterings dissimilarity and the types of possible dissimilarities that can exist among two clusterings in the context of community detection. We perform extensive evaluation to study the behaviour of the different similarity indices as a function of the dissimilarity type with both disjoint and non-disjoint clusterings. We finally provide practitioners with some insights on which similarity indices to use for the task at hand and how to interpret their values.

The rest of this paper is organized as follows. We propose our taxonomy in section II. We elaborate on the definition of dissimilarity in both disjoint, and non-disjoint clusterings in section III. We perform our evaluation and list our observations in section IV. The observations are discussed in section V, then we conclude our findings in section VI

II. A Taxonomy of Similarity Indices Used in Evaluating Community Detection Solutions

Her we provide a taxonomy (Figure 1) of similarity indices, i.e the metrics commonly used to quantify similarity among clusterings identified by community detection methods. The proposed taxonomy is constituted of three levels of comparison among clustering similarity indices. The top-level distinction is to specify whether the similarity index is used for comparing disjoint clustering where clusters do not overlap (we will refer to these indices as disjoint indices), or for comparing non-disjoint clustering where clusters might overlap (we refer to these indices as non-disjoint indices).

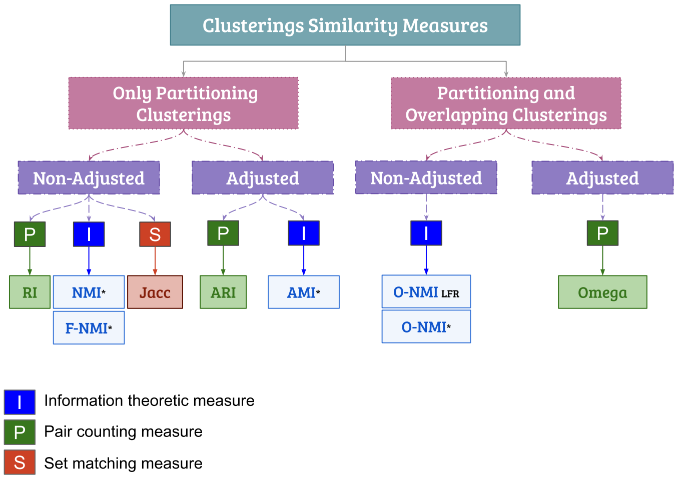


Fig. 1. A taxonomy of the similarity measures commonly used to evaluate community detection solutions. * refers to different ways of normalization (max, min, sqrt, mean, and joint)

The second level focuses on weather or not the similarity index is adjusted to take into consideration the by-chance agreement that can happen even between two randomly generated clusterings over the same node-set. The general formula used for adjustment is suggested by (5) to be:

$$Adjusted\_Index = \frac{Index\_Value - Expected\_Value}{Max\_Value - Expected\_Value} \tag{1}$$

Where $Index\_Value$ refers to the non-adjusted similarity score calculated by the index, Max_Value refers to the upper bound of the non-adjusted similarity score of that index, and $Expected\_index$ is a quantification for the by-chance agreement.

The third level in our taxonomy is to differentiate among three different techniques used in the literature to calculate similarity among clusterings, namely **(i)** pair counting which considers similarity on the level of node-pairs, i.e it counts the existence of two nodes together the same number of times in both clusterings as an agreement and then define similarity as an average of these agreements over all the possible ones, **(ii)** set matching which considers similarity on the set-level meaning that it treats communities of within each of the two compared clusterings as sets, and it define similarity as an average of the similarity among sets of the first clusterings with those of the other one, and **(iii)** mutual information which is borrowed measures from information theory. It treats each clustering $\mathcal{C}$ as a random variable X such that X is a vector of community memberships for each of the nodes in the considered node-set, then it maps the similarity between two clusterings $\mathcal{C}$,$\mathcal{C}'$ into calculating the mutual information among their relevant node membership vectors I($X_{\mathcal{C}}$,$Y_{\mathcal{C}'}$), which is a quantification in information theory of the extent to which knowing a random variable $X_{\mathcal{C}}$ contribute in reducing the entropy, uncertainty, of the another one $Y_{\mathcal{C}'}$.

The leaves in our taxonomy tree are the chosen similarity indices for our evaluation. For the sake of this paper, we choose to focus only the clustering similarity indices that are commonly used in the literature to evaluate similarity between community detection solutions (or between a community detection solution and a given ground truth). For further details about the mathematical formulations of these indices, we refer the reader to the original papers in which these indices were proposed. As shown in Figure 1, we chose to consider the following disjoint indices, Rand Index RI (6), Adjusted Rand Index ARI (5), Normalized Mutual Information NMI and Adjusted Mutual Information AMI (7), Fair Normalized Mutual Information F-NMI which is a version of NMI that penalizes the NMI score when differences in the number of communities exist among the compared clusterings (8), and Jaccard index (9), which is mainly meant to compare two sets but not two clusterings. We implemented a clustering-level Jacquard index by averaging the Jacquard coefficient of each community in a clustering when these communities are compared with the community in the ground truth with the highest intersection. As regards the non-disjoint indices, we chose to consider O-NMI$_{LFK}$ (10), O-NMI with the different normalization options proposed by (11) (i.e, Min, Max, Mean, and Sqrt), and Omega index (12) which is an extension of Rand Index for non-disjoint clusterings.

III. Types of Dissimilarity Among Clusterings

When it comes to comparing two disjoint clusterings $\mathcal{G}$ and $\mathcal{C}$, we claim that five main types of disagreements are possible to exist among them. Those are: **(i)** *misplaced nodes disagreement*, i.e a fraction of nodes that are not placed in the right community in $\mathcal{C}$ according to $\mathcal{G}$, **(ii)** *missing nodes disagreement*, i.e some nodes that do not exist in any community in $\mathcal{C}$ but they do in $\mathcal{G}$, **(iii)** *merging disagreement*, that is when a fraction of communities in $\mathcal{C}$ are constituted of some communities from $\mathcal{G}$ that are merged together, **(iv)** *splitting disagreement*, that is when a fraction of communities in $\mathcal{G}$ are split into multiple communities in $\mathcal{C}$, and **(v)** *random disagreement*, that is when a random combination of the aforementioned disagreements happen between $\mathcal{G}$ and $\mathcal{C}$.

To identify dissimilarity on the level of non-disjoint clusterings, we refer to two types of overlapping mentioned in (13), namely crisp overlapping where each node belongs fully to each community of which it is a member, and fuzzy overlapping where each node belongs to each community with a different probability. For the sake of this paper, we chose to focus only on the first type of overlapping (crisp overlapping). On a structural level, crisp overlapping can happen in various ways. It can be **(i)** hierarchical where the overlapping is represented by the existence of bigger communities that are constituted of other smaller communities merged together ,**(ii)** non-hierarchical where the overlapping is represented by either partial intersections among different communities, or simply by replicated communities within the clustering or a mix of both partial intersection and replicas, or **(iii)** mixed overlapping where both hierarchical and non-hierarchical overlapping exit. Figure 2 summarizes the different types of overlapping in non-disjoint clusterings.

Given the different types of overlapping discussed above, we still can identify only one type of non-disjoint dissimilarity, that is when the communities in a clustering $\mathcal{C}$ overlap differently compared to those in $\mathcal{G}$. To consider the different types of overlapping, we chose to have a separate experiment for each in our experiments as will be shown later.

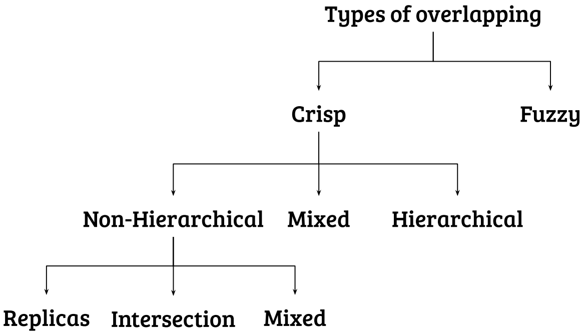


Fig. 2. Types of overlapping in non-disjoint clusterings

IV. Experiments

The main goal of our experiments is to study the behaviour of the different similarity indices used to compare community detection, clustering, solutions in different conditions. More specifically, we are interested in answering the following questions:

Q1 Do the different clustering similarity indices have the same trends in any condition and can they be, as a result, used interchangeably to compare community detection solutions?

Q2 Are the different clustering similarity indices equally sensitive to the different types of disagreements possible to happen between two clusterings?

Q3 Are the different clustering similarity indices on an agreement regarding what is considered dissimilar clusterings?

Q4 Given that these indices are normalized, is the relationship between the similarity score of each of these indices and the amount of disagreement between the compared clusterings linear?

To answer these questions, two main stages of evaluation were devised: one for similarity indices used to compare disjoint clusterings (section IV-A), and another for comparing similarity indices used with non-disjoint clusterings (sectionIV-B). We report our observations in Section IV-C.

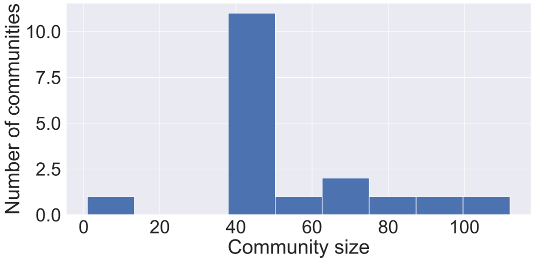


Fig. 3. Community size distribution for the clustering used as a ground truth throughout the experiments in this paper (unless mentioned otherwise)

We create a reference disjoint clustering, a ground truth $\mathcal{G}$, with respect to which we compare other disjoint clusterings, by detecting communities, using Louvain method (14), on a LFR benchmark graph (15) with following parameters: number of nodes n = 1000, power law exponent for the degree distribution = 2.6, power law exponent for the community size distribution =1.7, minimum degree = 2, maximum degree =n, minimum size of communities = 40, and maximum size of communities = n (using *networkx* Python package). This resulted on a clustering with 18 communities of which the community size distribution is illustrated in Figure 3. To have a non-disjoint ground truth $\mathcal{G}'$ with respect to which we compare other non-disjoint clusterings, we use the disjoint ground truth $\mathcal{G}$ to construct $\mathcal{G}'$ as will be explained in each experiment in section IV-B.

A. Comparing disjoint clusterings

To study the behaviour of similarity indices used to compare disjoint clusterings with respect to each type of the disjoing dissimilarities discussed in Section III, we construct a separate experiment for each. Each of these experiments starts with comparing two identical clusterings $\mathcal{G}$, $\mathcal{C}$, then at each step a new different disjoint clustering $\mathcal{C}_i$ is created and compared with $\mathcal{G}$ using the different disjoint similarity indices, namely NMI, AMI, F-NMI, RI, ARI, Jaccard. To improve the readability of the resulted figures in these experiments, we use only one type of normalization , MAX as inferred from (16), with the information theoretic indices (NMI, AMI, F-NMI). The non-disjoint clustering for each experiment were constructed for each experiment as follows:

- For the *misplaced nodes disagreement* experiment , a clustering $\mathcal{C}_i$ was derived from the ground truth $\mathcal{G}$ by choosing i nodes at random from $\mathcal{G}$ at each step i and moving them from their original communities in $\mathcal{G}$ to another randomly chosen community. Figure 4 reports the similarity scores of these disjoint clusterings at each step when compared with the ground truth $\mathcal{G}$.
- As to the *splitting disagreement* experiment (Figure 5), starting from the ground truth $\mathcal{G}$ communities, a new clustering $\mathcal{C}_i$ was generated by splitting one community

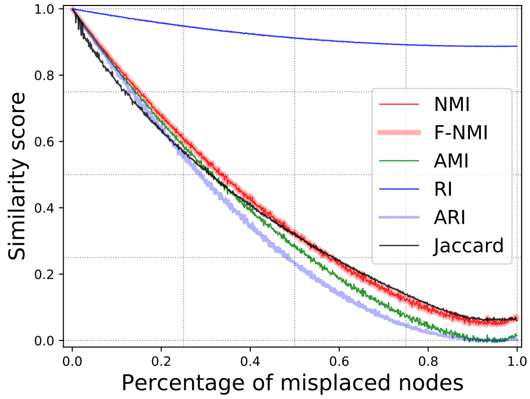


Fig. 4. Effect of misplaced nodes on disjoint similarity measures

from $\mathcal{G}$ chosen at random into two communities whose sizes are also chosen at random. Then, at each step i, a new clustering is generated by splitting a randomly chosen non-singleton community from the clustering at the previous step $\mathcal{C}_{i-1}$.

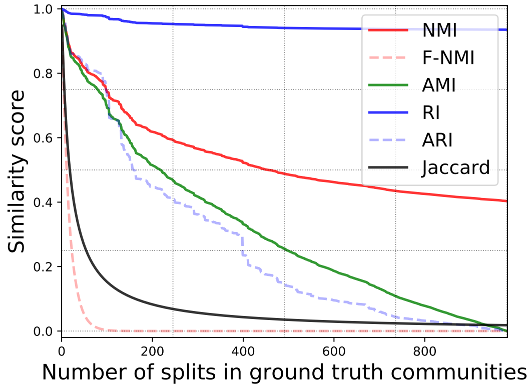


Fig. 5. Effect of a splitting disagreement on disjoint similarity measures

- As regards the *merging disagreement* experiment (Figure 6), starting from the ground truth $\mathcal{G}$ communities, a new ground truth clustering $\mathcal{G}'$ is derived from $\mathcal{G}$ by performing a number of random splits (100 splits). The goal from this step is to increase the number of communities in the ground truth so we allow for more merging possibilities and we can clearly see the effect of merging on the similarity indices. Starting from $\mathcal{G}'$, a new clustering $\mathcal{C}_i$ was generated by merging two different communities from $\mathcal{G}'$ chosen at random. Then at each step, a new clustering is generated by merging two different communities chosen at random from the clustering at the previous step $\mathcal{C}_{i-1}$ until a clustering $\mathcal{C}_{final}$ constituted of two communities is reached in the last step.
- For the *missing nodes disagreement* experiment (Figure 7), a clustering $\mathcal{C}_i$ was derived from the ground truth $\mathcal{G}$ by choosing i nodes at random from $\mathcal{G}$ at each step i, and removing them from their original communities. Since information theoretic indices do not account for missing nodes by definition and they require the two input clusterings node sets to be of the same size ($X_{\mathcal{C}}$, $X_{\mathcal{G}}$) so that a community membership is assigned for each node in $\mathcal{G}$ and $\mathcal{C}$, we assign a unique community membership in $X_{\mathcal{C}}$ for each of the missing nodes instead of totally removing the node.

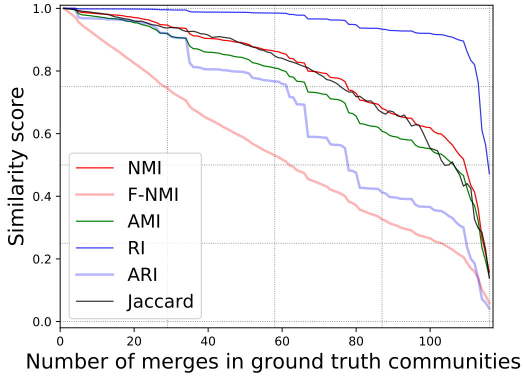


Fig. 6. Effect of a merging disagreement on disjoint similarity measures

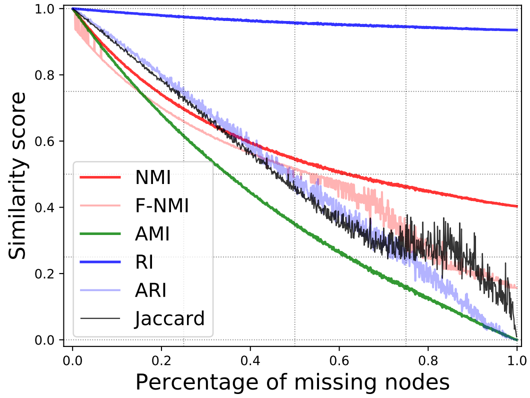
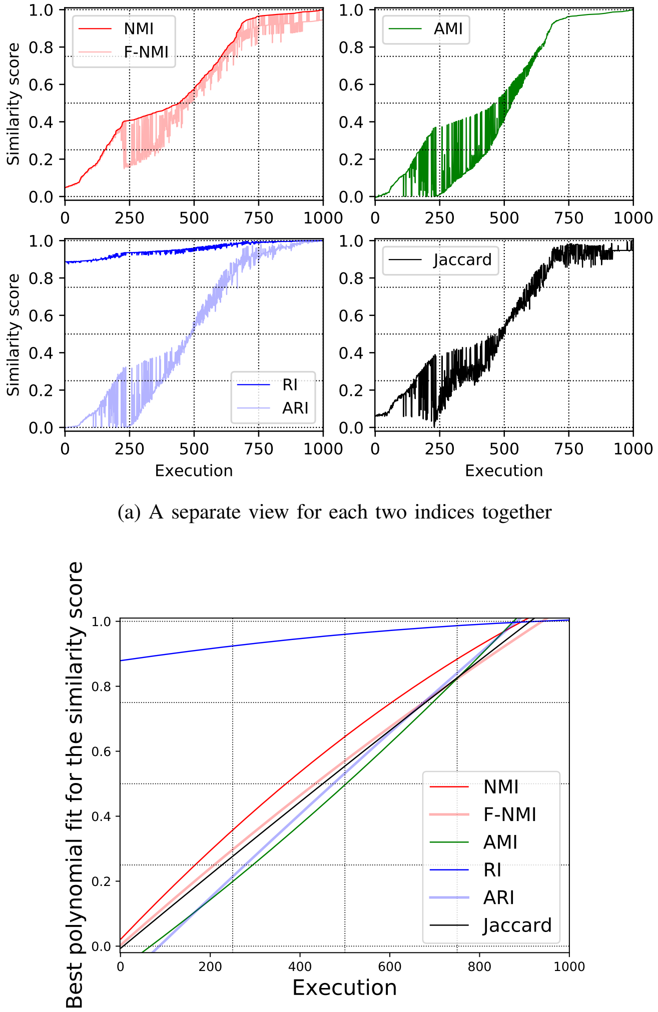


Fig. 7. Effect of missing nodes on disjoint similarity measures

- For the *random disagreement* experiment(Figure IV-A), a clustering $\mathcal{C}_i$ is generated at each step i by applying a sequence of up to four (the number is chosen at random between 1 and 4) disagreements of misplacing or missing nodes, merging or splitting communities on the clustering of the previous step $\mathcal{C}_{i-1}$. The goal of this experiment is to generally see if non-disjoint indices have similar trends when the clustering is constituted of random sequence of disagreements with respect to the ground truth (which might be the case in most real-world scenarios). Since applying a sequence of random disagreement at each step in this experiment does not always guarantee creating a clustering $\mathcal{C}_i$ that has more disagreement with the ground truth $\mathcal{G}$, the scores of all the indices are then ordered based on the ascending order of the NMI scores for visualization reasons. To improve the readability of this visualization, we split the figure into separate figures (Figure 8a) , then we show the best second order polynomial fit for all these indices (Figure 8b).



(a) A separate view for each two indices together

(b) The best second order polynomial fit of the disjoint indices visualized in Figure 8a

Fig. 8. main caption

B. Non-Disjoint clusterings

To consider the different types of overlappings mentioned in section III on different levels, we devise five different experiments to study the behaviour of non-disjoint similarity indices in different conditions. For these experiments, we create a non-disjoint ground truth $\mathcal{G}'$ starting from the disjoint ground truth clustering $\mathcal{G}$ used before, by randomly creating overlappings among the disjoint communities. We remind our reader that this step is not crucial for the validity of the non-disjoint indices as they can also be used to compare disjoint clusterings against non-disjoint ones, but we chose to do it to for the consistency of our experiments. The experiments are:

- *Pair-of-nodes level* overlapping. The goal of this experiment is to observe weather the pair-of-node level overlapping affect pair-counting indices (i.e Omega) differently than information theoretic ones (i.e O-NMI* and O-NMI$_{LFK}$). Starting from the ground truth clustering $\mathcal{G}'$, a clustering $\mathcal{C}_i$ is created in each step i, by randomly choosing a pair of nodes from a randomly chosen community in $\mathcal{C}_{i-i}$ and adding them together to another community so an overlapping is created on the level of pair-of-nodes. Figure 9 reports the similarity scores between $\mathcal{C}_i$, $\mathcal{G}'$ in each step.

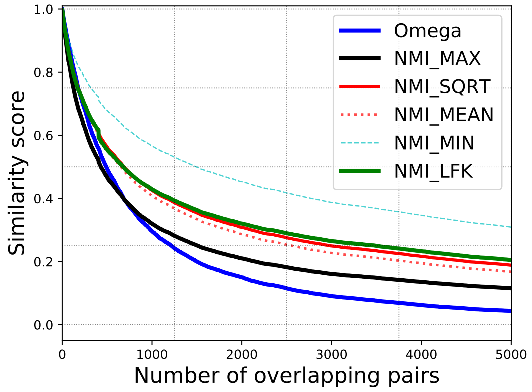


Fig. 9. Effect of pair-of-node level overlapping on non-disjoint similarity indices

- *Set-level* overlapping. The goal of this experiment is to observe weather the set-level overlapping affect pair-counting indices (i.e Omega) differently than information theoretic ones (i.e O-NMI* and O-NMI$_{LFK}$). Starting from the ground truth clustering $\mathcal{G}'$, a clustering $\mathcal{C}_i$ is created in each step i, by randomly choosing two different communities from $\mathcal{C}_{i-1}$ and creating a random intersection among them. Similarity scores among $\mathcal{C}_i$ and $\mathcal{G}'$ in each step are reported in Figure 10.

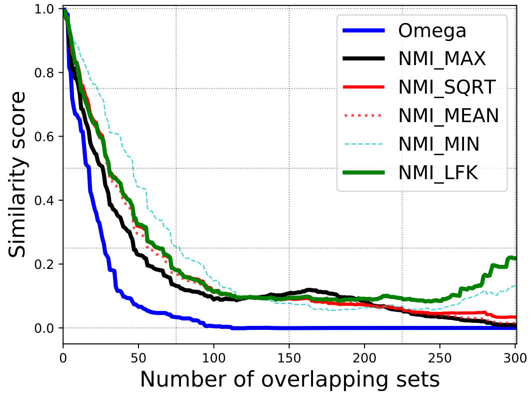


Fig. 10. Effect of set-level level overlapping on non-disjoint similarity indices

- *Hierarchical* overlapping. The goal of this experiment is to study the effect of hierarchical overlapping on non-disjoint similarity indices. We start from $\mathcal{G}'$, and at each step i, a clustering $\mathcal{C}_i$ is generated by adding a new community to $\mathcal{C}_{i-1}$ constituted of two randomly chosen communities from $\mathcal{C}_{i-1}$. Similarity scores are reported in Figure 11.

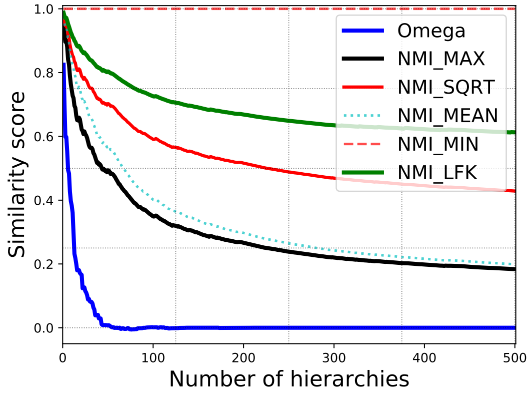


Fig. 11. Effect of hierarchical overlapping on non-disjoint similarity indices

- *Replication* overlapping. The goal of this experiment is to study the effect of replicated communities on non-disjoint similarity indices. We start from $\mathcal{G}'$, and at each step i, a clustering $\mathcal{C}_i$ is generated by simply replicating a randomly chosen community from $\mathcal{C}_{i-1}$. Similarity scores among $\mathcal{C}_i$, $\mathcal{G}'$ at each step are reported in Figure 12.

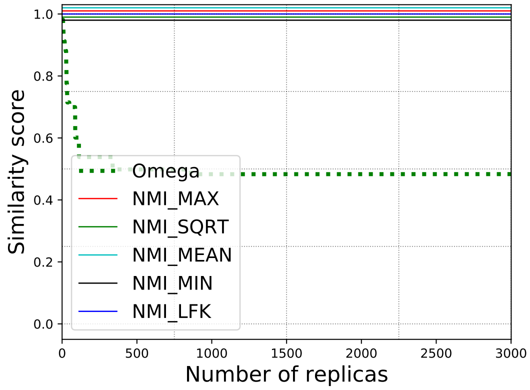


Fig. 12. Effect of replication overlapping on non-disjoint similarity indices

- *Random* overlapping. Here we study the effect of randomly chosen sequence of the aforementioned overlapping types on non-disjoint similarity indices. We start from $\mathcal{G}'$, and at each step i, a clustering $\mathcal{C}_i$ is generated by applying a sequence of up to four randomly selected types of overlapping on $\mathcal{C}_{i-1}$. The scores of all the indices are then ordered based on the ascending order of the Omega scores for visualization. Similarity scores among $\mathcal{C}_i$, $\mathcal{G}$ at each step are reported in Figure 13.

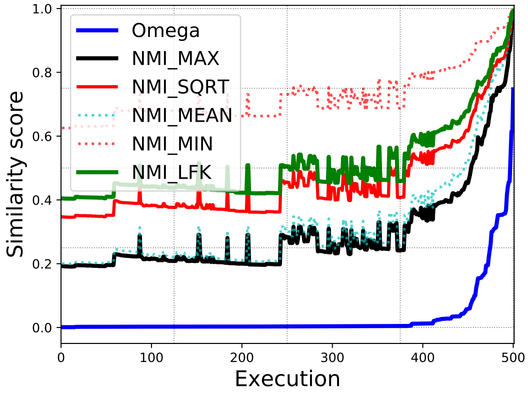


Fig. 13. Effect of random sequence of overlappings on non-disjoint similarity indices

C. Observations

A general observation on the disjoint indices is that they have different sensitivities with respect to the type of disagreement among the compared clusterings. Another observation is that rand index, in all cases, fails at quantifying dissimilarity among clusterings. Therefore, when we refer to disjoint indices in our observations, we mean to all the disjoint indices except rand index unless mentioned otherwise. Figure 4 suggests that while the adjusted rand index ARI proves to be a better extension of rand index RI to quantify the disagreement resulted by misplaced nodes, the adjusted mutual information AMI and the normalized mutual information NMI show a converging trend in this case. Moreover, it is important to notice that even though all the indices show a similar behaviour properly increasing the level of detected dissimilarity when a more and more nodes are misplaced, some do not hit 0 when the percentage of misplaced nodes is 100%. The figure shows also that 25% of misplaced nodes is enough to cause a drop to the half in the similarity scores suggesting a non linear behaviour between the similarity score and the amount of disagreement.

As to the effect of merging disagreement on the disjoint indices, Figure 6 reports that F-NMI is the most sensitive to the merging disagreement showing a quasi linear behaviour with respect to the number of merges in the compared clusterings $\mathcal{C}_i$ compared to the ground truth communities $\mathcal{G}$. The least sensitive is RI, followed by NMI and Jaccard, followed by AMI and then ARI. For NMI, jaccard and AMI, that show a similar behaviour, similarity still scores remarkably high (¿.8) when more than 50 merges have been performed. When it comes to the opposite type of disagreements,i.e splitting communities, Figure 5 shows a similar order in the sensitivities except for Jaccard, which seems to be more sensitive to splitting disagreements than to merging disagreements. In addition, splitting disagreement suggests that AMI is a better extension of NMI to identify this type of disagreements.

The last type of disagreement we have analyzed is the removal of nodes. In this case, ARI seems to have a linear behaviour (Figure 7). The same, approximately goes for Jaccard index, and AMI. Both NMI and F-NMI are less sensitive regarding missing nodes and it should be noticed how they do not hit 0 when the percentage of missing nodes in the clustering $\mathcal{C}_i$ is 100% with respect to the ground truth. That is because to the missing nodes are seen as isolated communities (instead of being totally removed).

The random disagreement experiment among disjoint clustering (Figure IV-A shows generally similar trends among disjoint indices when used to compare a ground truth $\mathcal{G}$ and a clustering $\mathcal{C}_i$ derived from the ground truth by applying a sequence of different types of disagreements. In most real-world scenarios, the identified clusterings can be constituted of random sequence of all the aforementioned disagreements with respect to a ground truth. In this case, non-disjoint indices can be interchangeably used if the goal is to compare the level of similarity among different pairs of clusterings such that the output is which pairs of clusterings are more similar that others. However, it seems still that different indices have different considerations about what is considered dissimilar.

As to the non-disjoint indices, a general observation is that the ranking of sensitivity seems to be stable among the different variations of O-NMI (with O-NMI$_{MAX}$ being the most sensitive among O-NMI variations) and with Omega being always the most sensitive among all.

While Figures 9 shows similar trends among non-disjoint indices with different sensitivities, 10 reports a non-intuitive behaviour for both O-NMI$_{MIN}$, and O-NMI$_{LFK}$ when they start to increase again after a certain amount of set-level overlapping while the other indices keep decreasing. The explanation for this has been discussed by (11) and that is when some hierarchies starts as a result of the random sets overlapping.

Figure 11 shows that O-NMI$_{MIN}$ score is not affected at all when the type of overlapping is hierarchical.

As to the replication overlapping, Figure 12 reports a 0 sensitivity with all O-NMI variations regarding this type of overlapping. While Omega index seems to be the only index that can be sensitive to this type of overlapping, still the level of sensitivity is the least when compared to Omega index sensitivity in other cases.

Even though the random overlapping experiment among non-disjoint clustering (Figure 13 shows generally similar trends among disjoint indices when used to compare a ground truth $\mathcal{G}$ and a clustering $\mathcal{C}_i$ derived from the ground truth by applying a sequence of different types of overlappings, we can not claim that all non-disjoint indices can be interchangeably used to compare non-disjoint clusterings given that they do not seem to have convergent scores. Based on the previous experiments we think that Omega index can be the most informative in this case, and O-NMI$_{MAX}$ can be a good compromise when the clustetings are big enough that the pair-counting mechanism performance becomes expensive.

V. Discussion

As mentioned in Section IV, the main goal of our experiments is to study the behaviour of different similarity quantification approaches used to compare clusterings identified by community detection methods. More specifically, we are interested in observing whether the different similarity indices have the same trends (**Q1**), analyzing their sensitivities to different types of disagreements (**Q2**), verify whether or not they agree on what is considered dissimilar (**Q3**), and check if the relationship between these indices and the level of agreement is linear (**Q4**). The ultimate goal is to provide practitioners with some insights about which indices to use for the task at hand and how to interpret their values.

As regards (**Q1**), different experiments show that similarity indices are do not always follow the same trends and that largely depends on the type of disagreement between the ground truth and the predicted clustering. With non-disjoint indices for example, the relationship among NMI, F-NMI, AMI, and ARI is more evident when the type of disagreement is a misplace-nodes disagreement. The is less evident among non-disjoint indices based on our experiments, and Omega index proves to be the most sensitive regarding all types of overlapping with O-NMI$_{MAX}$ being a good compromise when the given clusterings are too big such that the pair-counting mechanism performance becomes expensive.

As regards (**Q2**), It is clear based on our experiments that different indices have different sensitivities regarding the type of disagreement. We claim that this feature can be used to better understand predicted clusterings with respect to the ground truth. For example, with the sensitivity of each index reported in our experiments as a function of the type of disagreement, one can look at the values of F-NMI, Jaccard ,ARI, and AMI together. If all these values are convergent, one might expect that the type of disagreement between the ground truth and the predicted clustering is largely misplaced nodes disagreement. If these values are not convergent, one might expect, depending on the different sensitivities of each index, whether the type of disagreement is a splitting or merging disagreement. The same analysis is trickier with non-disjoint indices, given that the order of sensitivity among them seem to be consistent in most experiments. This also answers (**Q4**) stating that the relationship between similarity indices and the amount of disagreement is not linear in general as they do have different sensitivities regarding the type of disagreement.

Whether different indices are on an agreement regarding what is considered dissimilar (**Q3**), it is evident that this not the case as they do not always hit 0 together even though they tend to have similar trends in most cases.

VI. Conclusions

In this paper, we discuss the meaning of dissimilar clusterings in the context of community detection and we evaluate different similarity quantification techniques used in the literature to evaluate clusterings resulted by community detection methods. We propose a taxonomy for the similarity indices based on their use with disjoint or non-disjoint clusterings, weather they adjusted for the by-chance agreement or not, and the technique they use to define similarity (pair counting, set matching and information theoretic approach). We define the different types of dissimilarities that can exist between two disjoint-clusterings, or non-disjoint clusterings. Accordingly, we provide an extensive evaluation of the different similarity indices in with each type of dissimilarity.

Our experiments show that similarity indices do not always have the same trends. They do not agree on what is considered dissimilar, and they are not linear with respect to the amount of disagreement. We report also different levels of sensitivity among similarity indices used to compare disjoint clusterings based on the type of disagreement, which can be used to have more insights about the predicted clustering with respect to a given ground truth. The same does not hold with similarity indices used to compare non-disjoint clusterings showing that Omega index is the most sensitive index with all levels of non-disjoint disagreements.

Acknowledgment

This research is supported by the VIRT-EU project funded by the European Union's Horizon 2020 research and innovation program under grant agreement No 727040

Acknowledgment

We thank Michele Coscia for useful discussions and his efforts in referring to the implementations of different clusterings similarity indices.

References

[1] S. Papadopoulos, Y. Kompatsiaris, A. Vakali, and P. Spyridonos, "Community detection in social media," *Data Mining and Knowledge Discovery*, vol. 24, no. 3, pp. 515–554, May 2012. [Online]. Available: https://doi.org/10.1007/s10618-011-0224-z

[2] "Memory in network flows and its effects on spreading author =."

[3] U. Gargi, W. Lu, V. S. Mirrokni, and S. Yoon, "Large-scale community detection on youtube for topic discovery and exploration." 01 2011.

[4] M. E. Dickison, M. Magnani, and L. Rossi, *Multilayer Social Networks*. Cambridge University Press, 2016.

[5] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, no. 1, pp. 193–218, Dec 1985.

[6] W. M. Rand, "Objective criteria for the evaluation of clustering methods," *Journal of the American Statistical Association*, vol. 66, no. 336, pp. 846–850, 1971.

[7] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance," *J. Mach. Learn. Res.*, vol. 11, pp. 2837–2854, Dec. 2010.

[8] A. Amelio and C. Pizzuti, "Is normalized mutual information a fair measure for comparing community detection methods?" 08 2015.

[9] C. B. Urbani, "A statistical table for the degree of coexistence between two species," *Oecologia*, vol. 44, no. 3, pp. 287–289, Jan 1979.

[10] A. Lancichinetti, S. Fortunato, and J. Kertész, "Detecting the overlapping and hierarchical community structure in complex networks," *New Journal of Physics*, vol. 11, no. 3, p. 033015, mar 2009.

[11] A. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," *CoRR*, 10 2011.

[12] L. Collins and C. Dent, "Omega: A general formulation of the Rand index of cluster recovery suitable for non-disjoint solutions," *Multivariate Behavioral Research - MULTIVARIATE BEHAV RES*, vol. 23, pp. 231–242, 04 1988.

[13] S. Gregory, "Fuzzy overlapping communities in networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2011, no. 02, p. P02017, feb 2011.

[14] V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, p. P10008, oct 2008.

[15] A. Lancichinetti, S. Fortunato, and F. Radicchi, "Benchmark graphs for testing community detection algorithms," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 78, p. 046110, 11 2008.

[16] N. X. Vinh, J. Epps, and J. Bailey, "Information theoretic measures for clusterings comparison: Is a correction for chance necessary?" in *Proceedings of the 26th Annual International Conference on Machine Learning*, ser. ICML '09. New York, NY, USA: ACM, 2009, pp. 1073–1080.