Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 178 (2020) 355–364

Procedia Computer Science

www.elsevier.com/locate/procedia

9th International Young Scientist Conference on Computational Science (YSC 2020)

Community detection in node-attributed social networks: How structure-attributes correlation affects clustering quality

Petr Chunaev*, Timofey Gradov and Klavdiya Bochenina

National Center for Cognitive Technologies, ITMO University, Saint Petersburg, 199034 Russia

Abstract

The majority of parametric community detection (CD) methods working with node-attributed social networks (ASNs) focus on proposing new techniques and rarely pay much attention on the general analysis how ASN properties affect the corresponding CD quality. However, the latter mostly determines the applicability of a CD method in practice. To fulfil the gap, we investigate CD quality dynamics for ASNs with different structure-attributes correlation. The structure-attributes fusion model under consideration is a weight-based one that interpolates between the so-called fixed and non-fixed topology cases and generalizes a wide class of known weight-based models. Within the model, we first theoretically study the influence of correlation on CD quality and secondly illustrate our conclusions on specially constructed synthetic ASNs. Further, we test our conclusions on original and modified real-world ASNs. Our calculations indicate that the presence of correlation noticeably affects CD quality and that the simultaneous usage of network structure and attributes is not always reasonable within the weight-based fusion model under consideration. This makes the common suggestion that "adding attributes to structure leads to better CD results" questionable in certain cases.

Keywords: community detection; attributed social network; clustering quality; weight-based fusion model.

1. Introduction

Community detection (CD) in node-attributed social networks (ASNs) is a fast-growing area of social network analysis where multiple models and techniques have been appeared during the last decade [4, 5]. While classical CD approaches deal only with the structure or only with the attributes of a social network, CD methods for ASNs traditionally exploit the idea that node attributes, i.e. rich accompanying features of social actors e.g. age, gender, interests, added to the network structure may clarify and enrich the knowledge about the formation of communities

* Corresponding author. Tel.: +7-812-909-31-56.
E-mail address: chunaev@itmo.ru

10.1016/j.procs.2020.11.037

in the ASN [4, 5]. This idea is usually explained via the well-known homophily principle stating that like-minded social actors have a higher probability to be connected [13]. Furthermore, social science founding [11] suggests that node attributes can reflect and affect the community structure of a social network. As we have noticed, developers of CD methods for ASNs however partly misinterpret this concept as "adding attributes to structure leads to better CD results" that explains their focus on the technical side of the problem: adapt a technique to fuse available structure and attributes in the context of CD and show on particular ASNs that the new fusion technique produces (usually slightly) better CD results. The variety of the techniques and the corresponding problems are described in [5]. However, besides the technique used, theoretical basis for the necessity of structure-attributes fusion should be definitely taken into account. One should also understand how a particular fusion affects CD results and decide if the fusion is reasonable for ASNs with certain properties. Generally speaking, the latter mostly determines the applicability of a CD method in practice.

From the other side, CD theory for ASNs also suggests to extract a proper subspace of attributes in order to alleviate the *curse of dimensionality* [4]. The subspace-based methods use the *feature selection* paradigm particularly aiming at choosing attributes that are "relevant" to or "tightly correlated" with the network structure [12, 19]. Sometimes the mismatch between structure and attributes is stated to "negatively affect CD quality" [19]. Moreover, the existence of structure-attributes correlation is thought to "offer a unique opportunity to improve the learning performance of various graph mining tasks" [12]. However if only the subspace of attributes that are "relevant" to the structure is chosen, then one in fact obtains two sources of information that mainly duplicate each other. Following the homophily principle, we would say that the attributes *explain* the structure then but the idea that using the two analogous sources positively affects the CD results seems to be contentious, in our opinion.

In this study, we are interested in a theoretical and experimental study of how ASN properties, the structure-attributes correlation above all, and fusion parameters affect the corresponding CD quality. For our study, we choose one of the most popular and natural models [5] — the weight-based one that uses attributes similarity to assign weights on edges and produces a weighted network (instead of an ASN) that can be further divided into communities by classical CD graph algorithms. The choice is dictated by that such models are widely used for ASN CD and are sometimes superior in CD quality to some other models [5] (for example, to distance-based ones [1, 14], to probabilistic model-based ones [1, 2, 20] and to NNMF-based ones [2]).

Let us mention that there are many different approaches even to these simple models (we will give a short overview in Section 3). Besides different choice of weighting functions and fusion parameters, weight-based fusion models have different views on the raw ASN structure. Namely, there are approaches [5] that consider the ASN structure (aka *topology*) as (a) *fixed* (F), i.e. weights are assigned only on the initial edges of the ASN, and (b) *non-fixed* (NF), i.e. additional edges may be added to the ASN during the fusion process. We are unaware of any comparative analysis of these two concepts. Moreover, the general CD quality dynamics produced by weight-based models has been rarely studied until recently, if not taking into account illustrative results for particular ASNs [9, 16, 20].

Motivated by all the above-mentioned problems, we present our study of the dependence between varying structure-attributes correlation and CD quality dynamics. To be precise, the contributions are as follows. We study a weight-based model that provides an explicit control of the components and interpolates between F- and NF-topology and moreover generalizes a wide class of weight-based models for CD in ASNs. Within this model, we theoretically study how the choice of topology concept and the presence of structure-attributes correlation in an ASN affect CD quality. Furthermore, we test our theoretical conclusions in experiments with synthetic ASNs and original/modified real-world ASNs. We also discuss the feature selection paradigm described above in the context of ASN CD quality.

2. Problem statement and model description

2.1. Community detection problem statement

As it is done usually, we represent an ASN via triple $G = (\mathcal{V}, \mathcal{E}, \mathcal{A})$, where $\mathcal{V} = \{v_i\}$ is the set of nodes representing social actors, $\mathcal{E} = \{e_{ij}\}$ the set of edges representing the existing relations between the actors (so that e_{ij} is an edge between nodes v_i and v_j), and $\mathcal{A}$ the set of *positive real-valued* attribute vectors $A(v_i) = \{a_k(v_i)\}$ associated with the nodes in $\mathcal{V}$ and describing their features. (Note that the attributes are chosen numerical for simplicity. If one deals with nominal or textual attributes, it is common to use one-hot encoding or embeddings techniques to obtain their numerical representation.) The pairs $(\mathcal{V},\mathcal{E})$ and $(\mathcal{V},\mathcal{A})$ are correspondingly called ASN *structure* and ASN *attributes*. Furthermore, by $\overline{G} = (\mathcal{V},\overline{\mathcal{E}})$ we denote the complete graph with the set of vertices $\mathcal{V}$ and the set of edges $\overline{\mathcal{E}}$. In what follows, $|\mathcal{V}|$ stands for the size of $\mathcal{V}$, $|\mathcal{E}|$ for the size of $\mathcal{E}$, d for the dimension of attribute vectors $A = \{a_k\}$, $a_k(v_i)$ for the kth attribute of node v_i and K for the number of detected communities.

By *community detection* (CD) in an ASN we mean unsupervised dividing $\mathcal{V}$ of G into K disjoint subsets (*clusters* or *communities*) $C_k \subset \mathcal{V}$, with $C = \{C_k\}_{k=1}^{K}$, such that $\mathcal{V} = \bigcup_{k=1}^{K} C_k$ and a certain balance between the following two properties is achieved: *structural closeness*, i.e. nodes within a community are more densely connected than nodes in different communities, and *attributive homogeneity*, i.e. nodes within a community have similar attributes, while those in different ones do not. The basis for these properties is discussed in [4, 5] in detail.

By *CD quality dynamics* we mean the movement of CD quality measure values that occurs both while changing fusion model parameters and ASN properties.

2.2. Model description

Now we describe the model whose corresponding CD quality dynamics is studied. The model converts G into the weighted graph $G_W = (\mathcal{V},\mathcal{E}_p,\mathcal{W})$ with the set of weights $\mathcal{W} = \{W(e_{ij};\alpha)\}$ on a set of edges $\mathcal{E}_p$ as follows:

$$W(e_{ij};\alpha) = \alpha W_\mu(e_{ij}) + (1-\alpha)W_\nu(e_{ij}), \quad e_{ij} \in \mathcal{E}_p, \tag{1}$$

$$W_\mu(e_{ij}) = \frac{\mu(e_{ij})}{\sum_{e_{ij}\in\mathcal{E}_p} \mu(e_{ij})}, \quad W_\nu(e_{ij}) = \frac{\nu(e_{ij})}{\sum_{e_{ij}\in\mathcal{E}_p} \nu(e_{ij})}, \tag{2}$$

where $\alpha \in [0,1]$ is the *fusion parameter* that controls the impact of the components, μ is a *structural weight* function and ν an *attributive similarity weight* function. The parameter $p \in [0,1]$ is responsible for the *proportion of new edges created within the NF-topology concept*. In particular, the case $p = 0$ in (1) and (2) corresponds to the model in [6] where the F-topology version of (1) and (2) is proposed and considered. The set of edges used for weighting and determined by p is denoted by $\mathcal{E}_p$ so that $\mathcal{E} = \mathcal{E}_0 \subset \mathcal{E}_p \subset \overline{\mathcal{E}}$. Note also that $\alpha = 0$ and $\alpha = 1$ in (1) give weights based on attributes only and on structure only, correspondingly.

It is worth mentioning that the model (1) generalizes a wide class of weight-based models for CD in ASNs. The corresponding details on how (1) and (2) relate to known weight-based models will be given in Section 3, but from now on the reader should have in mind that the model (1) provides proper balance between the structural and attributive components, assigns weights on a controlled amount of edges thus may interpolate between the F- and NF-topology, and produces many of known weight-based models for certain chosen μ, ν, α and p.

For our experiments, we will use rather standard weight functions μ and ν, namely,

$$\mu(e_{ij}) = \begin{cases} 1, & e_{ij} \in \mathcal{E}, \\ 0, & e_{ij} \in \overline{\mathcal{E}} \setminus \mathcal{E}, \end{cases} \qquad \nu(e_{ij}) = \frac{A(v_i)\cdot A(v_j)}{\|A(v_i)\|_2\|A(v_j)\|_2} \in [-1,1], \quad e_{ij} \in \overline{\mathcal{E}}. \tag{3}$$

Note that ν is Cosine Similarity. Recall that our attributes are positive real and thus $\nu(e_{ij}) \in [0,1]$.

Furthermore, we choose to perform CD in the resulting weighted graph G_W and find the partition C by Weighted Louvain [3] as one of the most scalable and widely used for weighted graph clustering.

Note that experiments in [6] suggest that the choice of particular μ, ν and clustering algorithm does not qualitatively change the picture of CD quality dynamics.

2.3. Community detection quality measures

Since we are motivated by practical CD tasks where ground truth is rarely available, we use structural and attributive quality measures for evaluating CD quality. Namely, we use the traditional measures discussed in [4, Section 4.1]. Note that we measure CD quality on G for the partition C.

As for structural closeness, we apply a de facto standard measure of *Modularity* $\in [-1, 1]$ for a partition C (high values indicate dense connections between the nodes within clusters but sparse between those in different clusters). As for attributive homogeneity, we use a slight modification of Entropy that measures the degree of disorder of attributes within communities. Namely, for the case of *binary d-dimensional attributes*, we introduce *Anti-Entropy* $\in [0, 1]$ of a partition C as follows

$$\textit{Anti-Entropy}(C) = 1 - \sum\nolimits_{C_k \in C} \frac{|C_k|}{|V|} H(C_k), \qquad H(C_k) = -\sum\nolimits_{j=1}^{d} \frac{\phi(p_{kj})}{d \ln 2}, \qquad \phi(x) = x \ln x + (1-x)\ln(1-x),$$

where p_{kj} is the proportion of nodes in the community C_k with the same value on jth attribute.

Note that the measures above are defined so that their *high values refer to good CD quality, while low values refer to bad CD quality*.

2.4. Structure-attributes correlation measures

Let W_μ and W_ν with the summation over $e_{ij} \in \overline{\mathcal{E}}$ in (2) be samples of two variables. It is clear from the form of (1) that if W_μ and W_ν strongly depend on each other linearly, i.e. $W_\nu(e_{ij})$ is well-approximated by $aW_\mu(e_{ij})$ with some a (due to translation, the constant term may be omitted), then the dependence on α may even disappear. Indeed, for a fixed regression coefficient a the resulting weight (1) becomes directly proportional to $W_\mu(e_{ij})$ for any α: $W(e_{ij}; \alpha) = (\alpha(1-a) + a)W_\mu(e_{ij})$, i.e. the model (1) uses in fact only one weight component. This motivates us to ask whether the components in (1) are linearly or somehow else correlated as this may vastly affect the corresponding CD quality.

A straightforward way to estimate the linear correlation of the variables W_μ and W_ν is *Pearson's correlation coefficient* $r = \mathrm{cov}(W_\mu, W_\nu)/(\sigma_{W_\mu}\sigma_{W_\nu}) \in [-1, 1]$, where cov and σ are the corresponding covariance and standard deviations of the samples W_μ and W_ν (over $e_{ij} \in \mathcal{E}_p$). From the theoretical point of view, r however requires normally distributed variables to be properly interpreted and this is not the case for (3). Nevertheless one still can use the well-known connection of r and the regression coefficient a. A more flexible tool that assesses how well the relationship between two variables can be described by a monotonic function is *Spearman's rank correlation coefficient* $r_s \in [-1, 1]$ that is defined as r but between the rank variables generated by W_μ and W_ν. We will calculate both r and r_s in our experiments.

3. Related work

An overview of weight-based fusion models for CD in ASNs can be found in [5]. There are representatives of both F- and NF-topology concepts, e.g. [1, 6–8, 14, 17, 21] use the F one, while [2, 9, 10, 20] — the NF one. What is more, different α can be used, e.g. [1, 7, 8, 17, 21] use $\alpha = 0$ so that the weights (1) are in fact determined only by the attributes. Approaches with manually chosen $\alpha \in [0, 1]$ in (1) include [6, 8–10, 14, 20]. Note that the choice of α in (1) is difficult [4, 5] and has not been studied in full generality yet. Partial discussions on it are in [2, 6, 7, 9, 20].

With respect to the models above, (1) produces them as particular cases for different weighting procedures, including the choice of μ, ν, α and p. As has been already mentioned, $p = 0$ in (1) and (2) gives the model in [6]. If one chooses ν to be the matching coefficient, $\alpha = 0$ and $p = 0$, then (1) leads to the model in [17]. Jaccard or Cosine Similarity as ν and proper NF-topology procedure make (1) analogous to the model in [20]. As for the general CD quality dynamics associated with (1), it is rarely analyzed, besides the illustrative examples e.g. in [9, 16, 20]. The pioneering results on the F-topology case has been only recently obtained in [6] and will be discussed below.

Additionally, we are unaware of any research considering the influence of ASN structure-attributes correlation on the CD quality, besides the above-mentioned [12, 19] that propose to "improve" CD quality by erasing the attributes that are not correlating with the ASN structure. In our opinion, the point of view in [12, 19] is however rather questionable and still requires proper justification. Note that r and r_s are ideologically close to Assortativity and other similar measures for ASNs [15, 18] whose influence on CD quality seems to be an open problem, too.

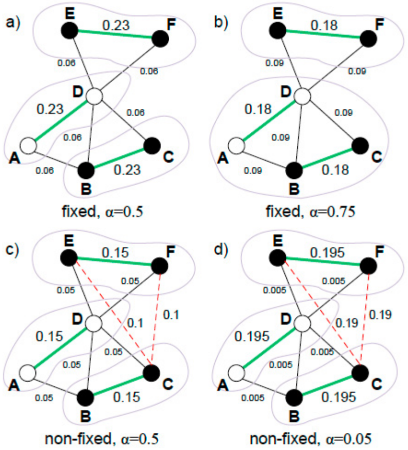


Fig. 1. Community detection: the difference between the F- and NF-topology cases.

4. Theoretical study

It it shown in [6] for synthetic ASNs with different intra/inter cluster densities that the quality dynamics of the F-topology weight-based model (i.e. with $p = 0$ in (1)) is non-linear although the model itself is linear. Furthermore, the model has the three regimes: (i) saturation of attributive measures when attributive homogeneity prevails ($\alpha \to 0$), (ii) saturation of structural measures when structural closeness prevails ($\alpha \to 1$), (iii) transition regime between (i) and (ii). A certain dependence of the length of the regimes on intra/inter cluster densities is observed for synthetic ASNs, too. It is also conjectured in [6] that the use of NF-topology would lead to widening the value range of attributive measures and keeping the value range of structural ones the same. It will be seen below that the regimes remain similar for the NF-topology case, i.e. for $p > 0$ in (1), but the conjecture from [6] is true only in part.

First note that for $p = 0$ the quality dynamics heavily relies on the raw structure of an ASN. Thus the F-topology case is generally unbalanced in the sense that the structure can prevail the attributes for some fusion parameters. As for the NF-topology case, adding new attributive edges to $\mathcal{E}$ (i.e. enlarging $p > 0$ and thus $\mathcal{E}_p$) potentially leads to better balance between the structure and the attributes. In this case structurally disconnected portions of a network have a chance to be connected via edges with high attributive weight. This implies that attributes are treated with more importance at the clusterization step than for $p = 0$. Under the normalization (2), some attributive weights can be higher than some structural ones. Thus, depending on α and p, the structure and the attributes can have equal impact to the CD results. Consequently, the observable value range of attributive measures can become wider and the values themselves can be higher than for $p = 0$ (this as conjectured in [6]). At the same time, once the attributes dominate for $\alpha \to 0$, structural weights may dissolve and structural measure values may become lower thus making the value range of structural measures wider (this is opposite to the conjecture in [6]). Recall that by construction $\alpha = 1$ is suggested to provide the highest values of structural measures and $\alpha = 0$ the highest values of attributive ones, on average.

Topology Conclusion. *For a fixed ASN, the NF-topology approach to the weight-based fusing model (1) usually results in widening the observable value range of the structural and attributive measures and the presence of lower and higher values of the structural and attributive measures, correspondingly, with respect to the F-topology approach.*

Fig. 1 illustrates the qualitative difference between the F- and NF-topology cases. We consider an ASN with binary attributes, $d = 1$, $|V| = 5$ and $|\mathcal{E}_0| = 8$, $|\mathcal{E}_p| = |\mathcal{E}_0|$ for Fig. 1a,b, $|\mathcal{E}_p| = 10$ for Fig. 1c,d. Structural weight μ is calculated according to (3), and attributive similarity weight ν is set to 1 if attributes are equal, and to 0 in the opposite case. Weighted Louvain algorithm with a default value of resolution parameter is used for CD. Under the F-topology, only two partitions are available. First one is observed e.g. for $\alpha = 0.5$ when attributive similarity dominates structural similarity but only for the initial set of edges $\mathcal{E}_0$. The increase of α for the F-topology leads to merging clusters $\{A, D\}$ and $\{B, C\}$ due to the increased importance of structural similarity. However, the subsets $\{E, F\}$ and $\{B, C\}$ will always be placed in different partitions even for small α. The NF-topology case (Fig. 1c,d) provides richer behaviour as it

Table 1. Synthetics networks with different structure-attributes correlation types.

Correlation type	Figure	r	r_s
Strongest +	–	1.00	1.00
Strong +	Fig. 2(a)	0.59	0.64
Zero	Fig. 2(b)	−0.07	−0.05
Strong –	Fig. 2(c)	−0.60	−0.62
Strongest –	Fig. 2(d)	−1.00	−1.00

can generate the same partitions as the F-topology as well as the new partition $\{A, D\}$, $\{B, C\}$, $\{E, F\}$. In the latter case, attributive similarity starts to dominate structural similarity even when nodes are not connected in the original ASN.

Below we discuss the influence of structure-attributes correlation on the CD quality dynamics. As follows from Section 2.4, CD quality dynamics for strong structure-attributes correlation is almost independent of α. In this case, the information components duplicate each other so it may be reasonable to use only one of them for CD, i.e. come back to the environment of classical CD methods based either on structure or attributes. Strong structure-attributes anti-correlation leads to the confrontation of reverse values of structural and attributive weights for different α in the NF-topology case, due to the fact that new connections are created only between nodes with similar attributes. By this reason, the influence of α should be noticeable and lead to wide value range of the quality measures between $\alpha = 0$ and $\alpha = 1$. However, in the F-topology case the initial network edges e_{ij} connect only nodes with attributes whose similarity is zero, i.e. $W_\nu(e_{ij}) = 0$. Thus, according to (1), the resulting weight is directly proportional to and thus determined only by the structural weight W_μ. Consequently, the dependence on α disappears. The corresponding constructive examples are given in Section 5.1. By continuity, the CD quality dynamics for ASNs with intermediate correlation rates is transitional between the above-mentioned extreme cases.

Correlation conclusion. *Within the weight-based fusion model (1), strong structure-attributes correlation in an ASN leads to weak dependence of the structural and attributive measures on α in both the F- and NF-topology cases. What is more, strong structure-attributes anti-correlation correspondingly leads to strong dependence of the structural and attributive measures on α in the NF-topology case, and to weak dependence on α in the F-topology case.*

5. Experimental study

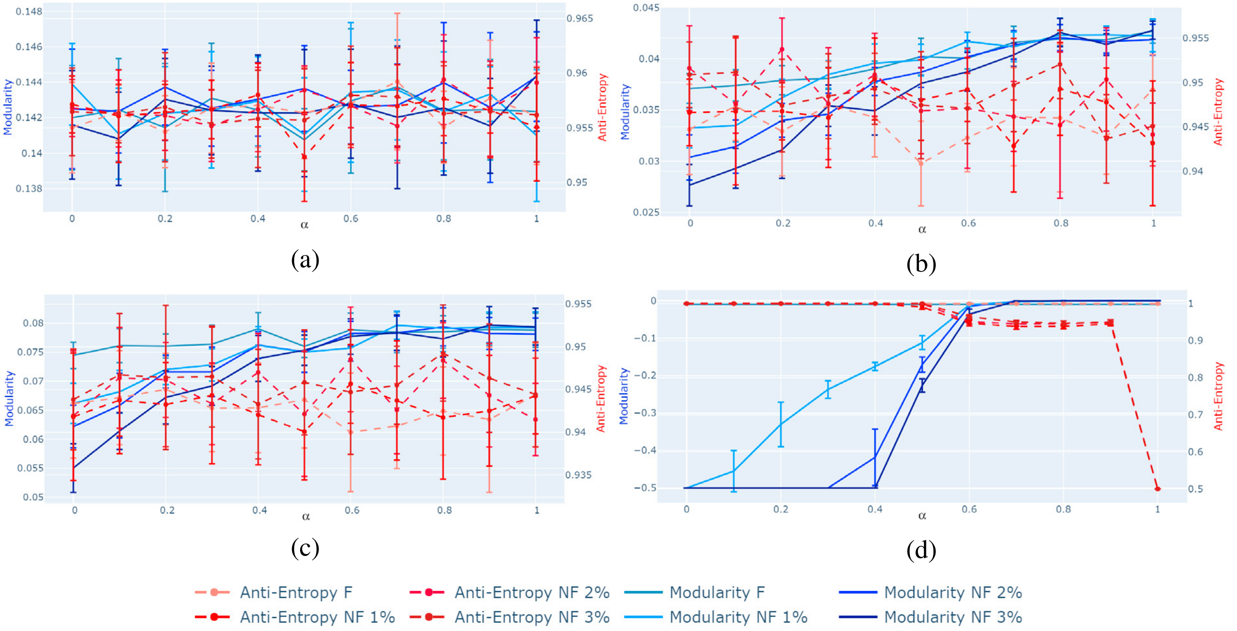
This section is devoted to illustrating[1] Topology and Correlation Conclusions on synthetic and real-world ASN. To produce the experimental study, we use the model (1) with the normalization (2), the weighting functions (3) and take α from 0 to 1 with step 0.1. Additionally, we embody the F- and NF-topology cases as follows. The F one corresponds to $p = 0$, i.e. the raw structure of a network is used to assign weights, $\mathcal{E}_0 = \mathcal{E}$. Within the NF one, we add edges between the current node and the ones with highest attributive similarity, if there was no initial structural edge. More precisely, in order to construct $\mathcal{E}_p$ we add attributive weights on a certain percent of the total amount of nodes in a synthetic ASNs: 1.00%, 2.00% and 3.00%, while the structural weight is zero for them. The corresponding percents for real-world ASNs are 0.25%, 0.50% and 0.75%.

As for the clusterization step, we use standard Weighted Louvain [3]. Its different runs may lead to different sets of communities and therefore we average the values of CD quality measures over 5 runs.

5.1. Experiments with synthetic networks

Let us start with the extreme cases of structure-attributes correlation $r = r_s = 1$ and anti-correlation $r = r_s = -1$, to illustrate Correlation Conclusion, see Table 1. Note that the presence of correlation is denoted in Table 1 by + and that of anti-correlation by −. It is worth mentioning that the densities of the synthetic ASNs chosen are almost equal.

[1] The source code and experimental results are presented on GitHub.



(a)

(b)

(c)

(d)

Fig. 2. Dynamics for the cases of different correlation: (a) $r = 0.59$, (b) $r = -0.07$, (c) $r = -0.60$, (d) $r = -1.00$.

The synthetic ASN with the *strongest correlation* is generated so that the nodes that share same attributes also share an edge, and vice versa. Namely, we first create two separate complete graphs with 100 nodes each and further merge them into the final graph (adding no edge between them). Attribute vectors in each connected component are generated to be orthogonal to attribute vectors in other components. In the general case when we have M separate complete components, the nodes in each mth cluster should have M-dimensional attribute vector with 1 as mth element and 0 in all others. Thus $\mu(e_{ij}) = \nu(e_{ij}) = 1$ for e_{ij} within each cluster and $\mu(e_{ij}) = \nu(e_{ij}) = 0$ for e_{ij} between different clusters. The experiments show that CD quality for this network is constant for all α and both the F- and NF-topology cases, namely, Modularity = 0.5 and Anti-Entropy = 1.0. Clearly, it satisfies Correlation Conclusion in the sense that there is no dependence of the structural and attributive measures on α for all the settings.

The latter analysis indicates that adding attributes to structure yields almost the same CD quality within the weight-based fusion model (1) in the case of strong correlation. Now imagine that we have an ASN with high-dimensional attribute vectors and apply the feature selection described in Introduction. As a result, we obtain a subset of attributes that are tightly correlated with the initial structure. If we perform the CD on the ASN with the new subset of attributes, we should expect the behaviour of quality measures stated in Correlation Conclusion and discussed in the previous paragraph. This actually makes the feature selection paradigm questionable as a tool for improving the CD quality on ASNs, with respect to the classical case when either structure or attributes are used within the CD process.

The network with the *strongest anti-correlation* is a bipartite graph, where the opposing subgraphs have inverse attributes. As seen from Fig. 2(d), the CD quality essentially depends on α in the NF-topology case as suggested by Correlation Conclusion. In the F-topology case there is no observable dependence on α.

Now r runs between the extreme cases, see Fig. 2(a)-(c) and Table 1. It is seen how the shape of the Modularity curve changes according to Correlation Conclusion. The dynamics of Anti-Entropy is not that clear due to the high standard deviations over runs, however the value range of both the measures becomes narrower when r grows.

The synthetic ASNs with intermediate correlation rates are generated as follows. Let n be the number of nodes (we choose $n = 100$) and k and m be the parameters controlling the correlation rate. Initially, binary one-dimensional attributes are generated randomly for the n nodes. Further, edges are added between the nodes according to the rule: k edges are created between each node and its top k attributes-similar ones and m edges between the node and top m attributes-dissimilar ones. By varying k and m one can achieve the correlation rates as in Fig. 2(a)-(c) and Table 1.

Table 2. Properties of the real-world ASNs under consideration.

Real-world ASN	Figure	r	r_s
Cornell	Fig. 3(a)	0.05	0.05
Cornell r-mod	Fig. 3(b)	0.22	0.17
Texas	cf. Fig. 3(a)	0.02	0.03
Washington	cf. Fig. 3(a)	0.05	0.06
Wisconsin	cf. Fig. 3(a)	0.05	0.06
Political Blogs	Fig. 3(c)	0.12	0.12
Bank Customers	Fig. 3(d)	0.01	0.01

Note that Fig. 2 also shows how the topology approach affects CD quality according to Topology Conclusion.

5.2. Experiments with real-world networks

We choose the following ASN datasets for the study of CD quality dynamics and its relation with the feature selection paradigm described in Introduction[2]:

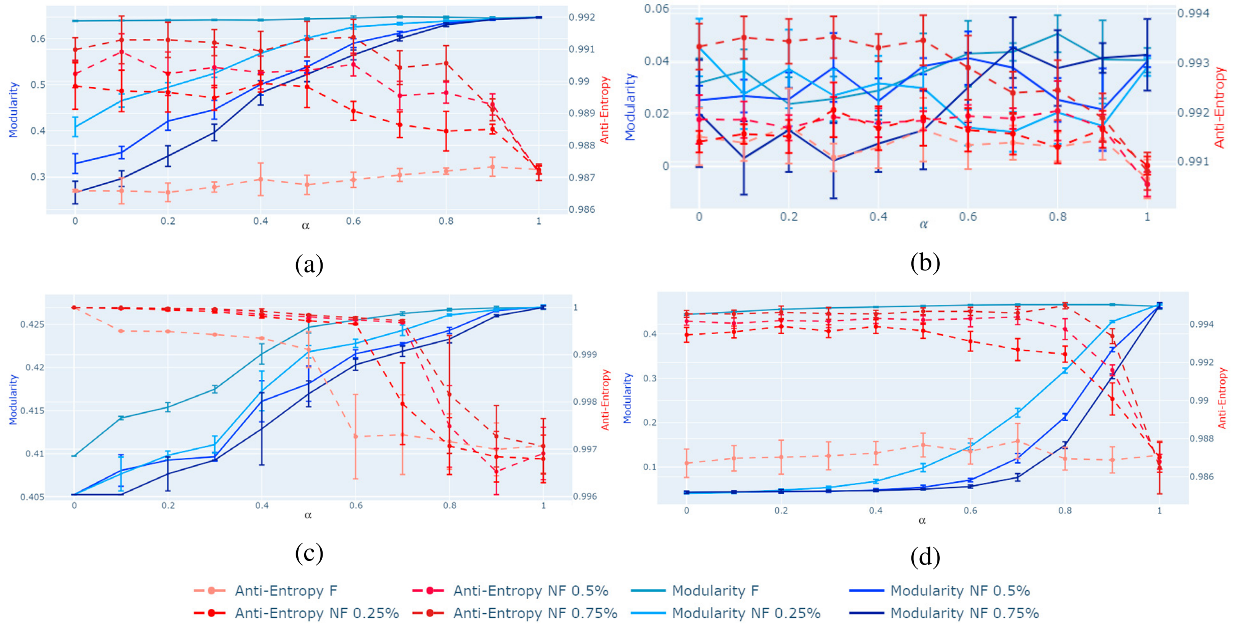
- *WebKB* is a collection of four ASNs (Cornell, Texas, Washington, and Wisconsin), totally of 877 web pages (nodes) with 1,608 hyperlinks (edges) gathered from four different universities Web sites. Each web page is associated with a binary vector whose elements indicate the the presence of a word from the vocabulary on that web page; the vocabulary consists of 1703 unique words.
- *Political Blogs* is an ASN of 1,490 webblogs (nodes) on US politics with 19,090 hyperlinks (edges) between these webblogs. Each node has a binary attribute describing its political leaning as either liberal or conservative.
- *Bank customers* is an ASN of 15,932 bank's group subscribers (nodes) with 200,639 connections between them (edges) and 19-dimensional binary node attributes representing subscriptions to different topical communities. The dataset is collected with anonymization from vk.com and contains a friendship network of subscribers of large regional bank community. According to bank's policy, this dataset cannot be made publicly available.

Results for all the networks in *WebKB* are similar so we present only those for Cornell, see Table 2 and Fig. 3(a).

First let us analyze Fig. 3(a) presenting the CD quality dynamics for the original Cornell network. It is seen that in the F-topology case the observable value range of the quality measures is quite narrow so that one can choose any $\alpha \in [0, 1]$ to get similar CD quality. In particular, $\alpha = 1$ provides the best quality when the F-topology concept is applied. However, according to Topology Conclusion, the value range substantially increases in the NF-topology case. One gets higher Anti-Entropy and lower Modularity values than in the F-topology case.

Now we modify the original Cornell network to demonstrate how structure-attribute correlation affects CD quality dynamics. We also discuss it in the context of the feature selection paradigm. Namely, we consider *Cornell r-mod*, the Cornell-based network with increased r and r_s, see Table 2. Initially we tried to extract a subset of attributes that are tightly correlated with the structure but it turned out that the highest structure-attribute correlation that we could achieve was $r = r_s = 0.12$. However, this is not enough for demonstration purposes. To overcome it, we erase edges that are not correlated with the attributes and add edges by attributive similarity in the initial Cornell network. This is done in a way to keep network density the same. The results for Cornell r-mod are in Fig. 3(b). It shows a rather weak dependence of the quality measures on α taking into account the narrower value range of the quality measures (six and two times less for Modularity and Anti-Entropy, correspondingly) although the resulting correlation coefficients are just $r = 0.22$ and $r_s = 0.17$. From the other point of view, if we applied the feature selection paradigm to Cornell r-mod and extracted the attributes tightly related with the structure, we could even obtain dynamics as in the strong

[2] Directed ASN WebKB and Political Blogs are simply considered as undirected.


(a)

(b)

(c)

(d)

Fig. 3. Dynamics for (a) original Cornell, (b) Cornell with modified correlation r, (c) Political Blogs, (d) Bank Customers.

correlation case where the CD quality is almost independent of α. Consequently, we could in fact use $\alpha = 0$ or $\alpha = 1$, i.e. either the structure or the attributes, to have very similar CD quality in terms of Modularity and Anti-Entropy.

As for *Political Blogs*, the dynamics of all the measures for both the F- and NF-topology cases is explicit, see Fig. 3(c). At the same time, the difference in value range of quality measures between the F- and NF-topology cases is notably less than that for WebKB. This might be the effect of correlation $r = 0.12$ in terms of Correlation Conclusion as this is the highest r among all the original real-world ASNs under consideration. Furthermore, $\alpha \approx 0.4$ providing the best balanced CD quality for the F-topology moves to $\alpha \approx 0.7$ for the NF one. Let us note that we also associate such an explicit dynamics for the F-topology with the dimension of attributes.

The dynamics for *Bank Customers* is revealing, see Fig. 3(d). First, it shows how dramatically the quality can change when moving from the F-topology to the NF one. Similarly to WebKB, the quality measure values for the F-topology are almost constant so one can choose any $\alpha \in [0, 1]$, especially the "classical case" ones $\alpha = 0$ or $\alpha = 1$, to have similar CD quality. However, once we add just 0.25% of new edges by attribute similarity, Modularity drops from 0.45 to 0.00 and Anti-Entropy jumps from 0.987 to 0.994. Interestingly, adding more edges does not change the values much. It can be observed that the best balanced CD quality is for $\alpha \approx 0.9$ in the NF-topology case.

6. Conclusions

In this study, we considered the generalized weight-based model (1) that provides an explicit control of the components and interpolates between the F- and NF-topology cases. In the context of the model, we first theoretically analyzed how the F- and NF-topology and ASN structure-attributes correlation affect CD quality dynamics. In particular, we showed that the NF-topology usually results in widening the value range of structural and attributive measures allowing one to obtain lower and higher values for the former and for the latter, correspondingly, with respect to the F-topology case. Moreover, we concluded that strong structure-attributes correlation in an ASN leads to weak dependence of the quality measures on α. What is more, we have confirmed our conclusions in experiments with special synthetic ASNs and several real-world ASNs. It was observed that a simultaneous usage of structure and attributes is hardly reasonable for ASNs with strong structure-attributes correlation. Furthermore, the examples with strong correlation showed that the feature selection paradigm aiming at extracting the subset of available attributes that are tightly correlated with the structure is sometimes questionable with respect to the improvement of CD quality within the weight-based fusion model (1). Overall, the above-mentioned facts indicate that the common suggestions that "adding attributes to structure leads to better CD results" and that "the presence of structure-attributes correlation improves the performance of CD tasks" are not universal recipes, at least within the model (1).

As for the possible future research, it is interesting to study how ASN properties (in particular, structure-attributes correlation) affect CD quality within other fusion models in [5]. From our side, we expect conclusions similar to ours.

Acknowledgements

This research is financially supported by Russian Science Foundation, Agreement 17-71-30029 with co-financing of Bank Saint Petersburg.

References

[1] Akbas, E., Zhao, P., 2019. Graph clustering based on attribute-aware graph embedding, in: Karampelas, P., Kawash, J., Özyer, T. (Eds.), From Security to Community Detection in Social Networking Platforms. Springer International Publishing, Cham, pp. 109–131. doi:10.1007/978-3-030-11286-8_5.

[2] Alinezhad, E., Teimourpour, B., Sepehri, M.M., Kargari, M., 2020. Community detection in attributed networks considering both structural and attribute similarities: two mathematical programming approaches. Neural Computing and Applications 32, 3203–3220. doi:10.1007/s00521-019-04064-5.

[3] Blondel, V.D., Guillaume, J.L., Lambiotte, R., Lefebvre, E., 2008. Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment 2008, P10008. doi:10.1088/1742-5468/2008/10/p10008.

[4] Bothorel, C., Cruz, J.D., Magnani, M., Micenková, B., 2015. Clustering attributed graphs: Models, measures and methods. Network Science 3, 408–444. doi:10.1017/nws.2015.9.

[5] Chunaev, P., 2020. Community detection in node-attributed social networks: A survey. Computer Science Review 37, 100286. doi:10.1016/j.cosrev.2020.100286.

[6] Chunaev, P., Nuzhdenko, I., Bochenina, K., 2019. Community detection in attributed social networks: A unified weight-based model and its regimes, in: 2019 International Conference on Data Mining Workshops (ICDMW), pp. 455–464. doi:10.1109/ICDMW.2019.00072.

[7] Combe, D., Largeron, C., Egyed-Zsigmond, E., Gery, M., 2012. Combining relations and text in scientific network clustering, in: Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), IEEE Computer Society, Washington, DC, USA. pp. 1248–1253. doi:10.1109/ASONAM.2012.215.

[8] Cruz Gomes, J.D., Bothorel, C., Poulet, F., 2011. Semantic clustering of social networks using points of view, in: CORIA: conférence en recherche d'information et applications 2011, Avignon, France. pp. 1–8. URL: https://hal.archives-ouvertes.fr/hal-00609291.

[9] Dang, T.A., Viennet, E., 2012. Community detection based on structural and attribute similarities, in: International Conference on Digital Society (ICDS), pp. 7–14.

[10] Jia, C., Li, Y., Carson, M.B., Wang, X., Yu, J., 2017. Node attribute-enhanced community detection in complex networks. Scientific Reports 7:2626, 1–15. doi:10.1038/s41598-017-02751-8.

[11] Kossinets, G., Watts, D.J., 2009. Origins of homophily in an evolving social network. American Journal of Sociology 115, 405–450. URL: http://www.journals.uchicago.edu/doi/abs/10.1086/599247.

[12] Li, J., Guo, R., Liu, C., Liu, H., 2019. Adaptive unsupervised feature selection on attributed networks, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, p. 92–100. doi:10.1145/3292500.3330856.

[13] McPherson, M., Smith-Lovin, L., Cook, J.M., 2001. Birds of a feather: Homophily in social networks. Annual Review of Sociology 27, 415–444. doi:10.1146/annurev.soc.27.1.415.

[14] Meng, F., Rui, X., Wang, Z., Xing, Y., Cao, L., 2018. Coupled node similarity learning for community detection in attributed networks. Entropy 20. doi:10.3390/e20060471.

[15] Mulders, D., de Bodt, C., Bjelland, J., Pentland, A., Verleysen, M., de Montjoye, Y.A., 2018. Inference of node attributes from social network assortativity. Neural Computing and Applications , 1–21.

[16] Nawaz, W., Khan, K.U., Lee, Y.K., Lee, S., 2015. Intra graph clustering using collaborative similarity measure. Distributed and Parallel Databases 33, 583–603. doi:10.1007/s10619-014-7170-x.

[17] Neville, J., Adler, M., Jensen, D., 2003. Clustering relational data using attribute and link information, in: In Proceedings of the Text Mining and Link Analysis Workshop, 18th International Joint Conference on Artificial Intelligence, pp. 9–15.

[18] Newman, M.E.J., 2003. Mixing patterns in networks. Phys. Rev. E 67, 026126. doi:10.1103/PhysRevE.67.026126.

[19] Qin, M., Jin, D., Lei, K., Gabrys, B., Musial-Gabrys, K., 2018. Adaptive community detection incorporating topology and content in social networks. Knowledge-Based Systems 161, 342 – 356. doi:10.1016/j.knosys.2018.07.037.

[20] Ruan, Y., Fuhry, D., Parthasarathy, S., 2013. Efficient community detection in large networks using content and links, in: Proceedings of the 22Nd International Conference on World Wide Web, ACM, New York, NY, USA. pp. 1089–1098. doi:10.1145/2488388.2488483.

[21] Steinhaeuser, K., Chawla, N.V., 2010. Identifying and evaluating community structure in complex networks. Pattern Recognition Letters 31, 413 – 421. doi:10.1016/j.patrec.2009.11.001.