2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

Multigraph transformation for community detection applied to financial services

Safa El Ayeb
Orange
Normandie Univ, UNICAEN
ENSICAEN, CNRS, GREYC
Caen, France
safa.elayeb@orange.com

Baptiste Hemery
Orange
Caen, France
baptiste.hemery@orange.com

Fabrice Jeanne
Orange
Caen, France
fabrice.jeanne@orange.com

Christophe Charrier
Normandie Univ, UNICAEN
ENSICAEN, CNRS, GREYC
Caen, France
christophe.charrier@unicaen.fr

Estelle Cherrier
Normandie Univ, UNICAEN
ENSICAEN, CNRS, GREYC
Caen, France
estelle.cherrier@ensicaen.fr

***Abstract*—Networks have provided a representation for a wide range of real systems, including communication networks, money transfer networks and biological systems. Communities represent fundamental structures for understanding the organization of real-world networks. Uncovering coherent groups in these networks is the goal of community detection. A community is a mesoscopic structure with nodes heavily connected in their groups by comparison to the nodes in other groups. Communities might also overlap as they may share one or multiple nodes. This paper lays the foundation for an application on transactional multigraphs (networks of financial transactions in which nodes can be linked with multiple edges), through the discovery of communities. Due to their complexity, our goal is to find the most effective way of simplifying multigraphs to weighted graphs, while preserving properties of the network. We tested five weights' calculation function and community detection algorithms were applied. A comparison of the outputs based on extrinsic and intrinsic evaluation metrics is then held.**

***Index Terms*—Social Network Analysis, Overlapping community detection, multigraph, graph transformation, Financial service.**

I. Introduction

The framework of networks provides an exceptional tool for the analysis of complex systems of interacting objects. Real networks decompose into densely connected modules, also called communities, with sparse ties between them. Communities correspond to behavioral or functional entities within the network. Despite the progress achieved in the area, community detection is still a complex task due to the lack of universal definition of community. Every proposed approach defines its own idea of a community related to the studied context [1], [2]. There have been multiple studies in the literature targeting community detection in social networks [3]. However, most of them focused on single-relational networks.

Motivated by an application around transactional data analysis, this paper studies graph transformation for overlapping community detection. Transactions' network is a directed multigraph, where nodes represent users, edges represent money transfers, and allowing multiple edges between a pair of nodes. In practice, multigraphs haven't been widely studied in comparison to simple graphs. Simple graphs can be derived from multigraphs, by collapsing the multiple edges between two nodes into one simple edge. However, this process discards information from the original graph [4].

In this paper, we propose to compare different multigraphs reduction operations based on different function of edges' weights calculation. Our goal is to understand to what extent the reduction of transactional multigraphs is beneficial to the overall community detection quality and which type of reduction is more efficient. In order to minimize the loss of information caused by the graph transformation, the proposed method consists in adding a weight on the simplified edge using available information about the transactions. We aim to generate different graphs, using different weighting functions. In what follows, we will use the terms node, vertex, vertice, and edge, link interchangeably.

The rest of the paper is organized as follows. Section II presents the context of the study and preliminary definitions about overlapping community detection and evaluation metrics. Section III illustrates the problem of multigraph reduction, and the contribution of this paper through the proposed method. In section V our methodology is explained as well as data deployed for the tests. Finally, in section VI tests results are presented and discussed.

II. Context

A. Use case around financial data

In this paper, we consider a graph consisting of financial transactions for a money transfer service where nodes repre-

The authors would like to thanks Orange and the ANRT for funding this work.

IEEE/ACM ASONAM 2022, November 10-13, 2022
978-1-6654-5661-6/22/$31.00

sent users and edges represent transactions. For such a service, transactions take place between customers for various reasons such as sending money for family, transfers between friends, salary payment, etc. Social network analysis has already been successfully used on banking and money transfer data to carry out socio-economic studies [5], uncover customer buying habits [6], fight fraud [7], etc.

The social network established from transactions data has two important features, namely link directedness and multiplicity what makes it a directed multigraph. More precisely, the graph is directed because each transaction involves a sender and a receiver. This network is a multigraph, given the fact that between any pair of users, there can be several transfers (edges in the graph) over a period of time. When these features are ignored for community detection, there might be issues over quality and accuracy of communities obtained [8]. Figure 1 displays a fraction of a transactional network.

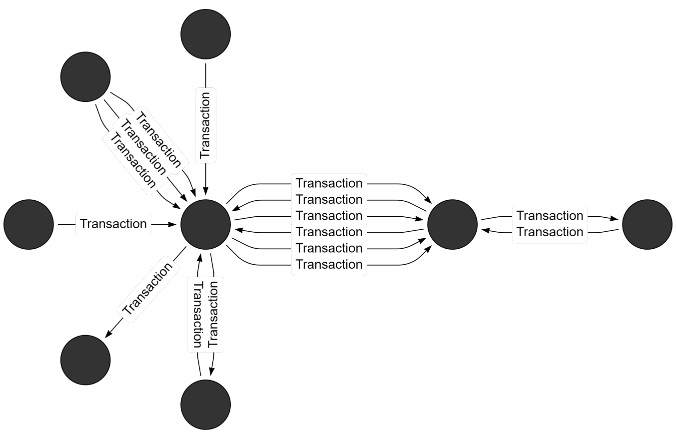

Fig. 1. Example of transactions' graph.

As depicted in figure 1, each pair of users (nodes) can have multiple transactions (edges) one way or the other. In the transactional graph, each edge is qualified by a set of attributes such as the transaction amount, the date, the status (success or failure), etc. In this work, we will rely on these attributes more particularly transaction amount and date in order to convert the multigraph on a weighted graph.

B. Overlapping communities detection

Although no formal consensus was reached on a definition that captures the gist of a community, it may be defined as a cohesive group where nodes interact with each other significantly more than they do with other nodes outside the group [3], [9]. In this work, we are particularly interested in overlapping communities. Overlapping community structure is a natural phenomenon observed in real networks. Unlike crisp communities, overlapping communities may share one or more nodes: a node can simultaneously be part of multiple communities of different scopes and levels, such as family, friends, work, city, etc. Therefore, in the last years, there has been a growing interest in identifying communities that are not necessarily disjoint [8], [10], [11].

Overlapping communities are studied in the literature in various contexts. In the context of e-commerce, [12] uses overlapping communities in order to prove the significance of overlapping nodes in viral marketing. They uncovered that these nodes are the most influential nodes. In the same framework, [13] studies reviews' spammers (who post fake reviews) based on overlapping community detection in the reviews graph. Overlapping nodes are identified as highly suspicious users. The authors have then developed a trust rank algorithm based on community detection results. Several tests have proven that their community rank algorithms outperformed other existing methods.

Finally, in the context of mobile social networks, Kim and Kim [8] studied overlapping communities in the network of phone calls and texts exchanged between users. They studied the detection of overlapping communities taking into consideration real social network features such as nodes overlapping, the weight and the direction of the edges, and the hierarchical structure of the network. For a complete study of overlapping community detection, we refer the reader to the works [11].

III. Problem statement

A. From Multigraphs to weighted graphs

As data complexity increases, the classic definition of a simple graph falls short to represent the complex semantics of real world networks. More specifically, in these networks we can observe multiple interactions describing the same or different types of relationships between nodes. For example, a group of users may have various communication channels like phone calls, emails, messaging, and so on, or have multiple exchanges of the same type. Graphs having several edges connecting the nodes are called multigraphs.

In this paper, we consider the case were nodes have multiple interactions of the same type. Most of the work regarding this specific type of graphs is focused on multi-types edges, or what is called multi-layers networks. However, multigraphs can represent multiple real systems, like co-authorship collaboration networks [14], online social networks [15], communication networks [16], etc. Despite the large amount of interest in multigraphs, relatively little attention has been paid to multigraphs with the same type of edges.

Although multigraphs exploitation in different applications such as community detection provide rich information to the analysis, their characteristics also intensify the complexity of calculation and processing. Therefore multigraph reduction or compression appears to be a solution to mitigate this scale-up. The reduction of a multigraph consists in reducing the number of edges or nodes existing in the network for a simpler representation.

While multigraph's reduction may lead to loss of information and details, it allows to eliminate the noise that can be present in the graph. We expect that the transition from a multigraph into weighted graph would allow to restore the information lost during the compression. Figure 2 shows an example of an undirected multigraph reduction to a simple undirected weighted graph.

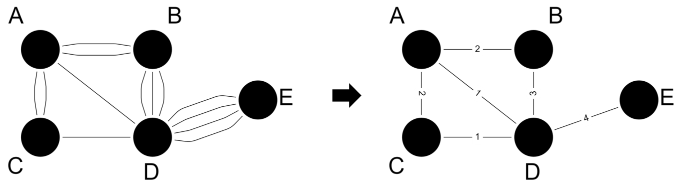

Fig. 2. Example of multigraph's reduction to simple weighted graph.

Weighted graphs have been studied in different contexts as many real world networks are intrinsically weighted. Their edges having different strengths depict stronger or weaker ties and flows between nodes. As any type of graphs, weighted graphs have attracted attention for different application, especially for community detection [8], [17], [18]. These studies have led to various algorithms for the detection of disjoint and overlapping communities.

B. Notations

A graph $G = (V, E)$ is a mathematical object, where $V = \{v_1, v_2, \ldots, v_n\}$ denotes the set of vertices and $E \subseteq V \times V$ the set of edges. We define the neighborhood of a node v_i of G, or $N(v_i)$, as the set of nodes adjacent to v_i. In a graph, an edge $e \in E(G)$ may be undirected, or directed from the sender node to the receiver node.
Let v_i and v_j be two vertices of G. We use $\mu(v_i, v_j)$ to denote the number of edges joining v_i and v_j in G.

Definition 1. A graph G is weighted, if it associates to every edge $e \in E$ a weighting function that gives a real non negative such that

$$W(e) \geq 0$$

If the weighted graph is undirected, then its weighting function is symmetric i.e. $w(v_i, v_j) = w(v_j, v_i)$.

Definition 2. The multiplicity of a vertex v_i is given by

$$\mu(v_i) = \max_{v_j \in V \setminus \{v_i\}} \mu(v_i, v_j)$$

The multiplicity of a graph G is defined by

$$\mu(G) = \max_{v_i \in V} \mu(v_i)$$

A graph is said to be simple if it has multiplicity at most 1. Otherwise, if a graph has multiple edges between any pair of its nodes, it is called a multigraph.

Definition 3. For a given number k, we define a cover as a family of k subsets of nodes $C = \{C_1, C_2, \ldots, C_k\}$ where each C_i is called a community.

The goal in community detection is to find a cover C which best describes the community structure of the graph. Namely, communities are sub-graphs of a graph.

In this work, we are interested in identifying overlapping communities, which can be defined as follows.

Definition 4. An overlapping community structure C can be defined as a cover of V in n communities $C = \{C_1, C_2, \ldots, C_n\}$ where a node v_i can participate in one or more communities C_i with a belonging factor α such that $0 \leq \alpha \leq 1$, and

$$\sum_{i}^{n} \alpha_{v_i} = 1, \forall v_i \in V, \forall C_i \in C$$

In section V-B, we will refer to a result and ground truth overlapping communities respectively as $R = R_1, R_2, \cdots, R_n$, and $G = G_1, G_2, \cdots, G_m$.

IV. Proposed method

Our goal in this work is to simplify the structure of the transactional graph by reducing the number of edges, while preserving the graph properties, and the strength of ties between nodes. Weights on edges will provide information about the relation between two nodes. They will also characterize the connection between these two nodes, based on the exchanges they had. In our studied graph, each edge e represent a financial transaction between a sender and a receiver. A transaction carries a set of characteristics such as the amount and the date which are converted into edge attributes. A couple of nodes or users v_i, v_j can engage in multiple transactions and thus be linked by multiple edges. Reducing the graph consists on combining these edges in a single edge with a weight $w(v_i, v_j)$ We expect that reducing the size of the multigraph would allow better understanding of the graph patterns, as well as provide a better and faster result for community detection algorithms. In the context of this paper, we chose five methods to transform the multigraph to an oriented weighted graph. For each method, the weight of the edge will be calculated differently. These methods are called according to the weighting functions: *occurrence*, *sum*, *mean*, *monthly mean* and *temporal score*.

For all five methods, we proceed as follows: all the directed edges from one sender node to a receiver node will be replaced by one directed weighted edge. For a couple of nodes in the graph, the weight scores of the resulted edges are calculated as follows :

a) Occurrence: : The resulting edge weight represents the number of edges between the two nodes before the simplification.

$$w(v_i, v_j) = \mu(v_i, v_j) \tag{1}$$

b) Sum: : The edge weight of the second method is equal to the sum of transactions amounts TA between the transmitting node v_i and the receptive node v_j over the study period;

$$w(v_i, v_j) = \sum_{i=1}^{\#(TA)} transaction\, amount(v_i, v_j) \tag{2}$$

where $\#(TA)$ is the number of transactions between v_i and v_j.

c) Mean: : The resulting edge weight is equal to the mean of the transactions amounts exchanged over the study period.

$$w(v_i, v_j) = \frac{\sum_{i=1}^{\#(TA)} transaction\,amount(v_i, v_j)}{\mu(v_i, v_j)} \quad (3)$$

d) Monthly mean: : The fourth method, similar to the two previous methods, is based on temporal mean of transactions over the study span.

$$w(v_i, v_j) = \frac{\sum_{i=1}^{\#(TA)} transaction\,amount(v_i, v_j)}{\sum months} \quad (4)$$

e) Temporal score: : The last method of weight calculation is based on a score mixing transactions amounts and temporal variables. The score is based on all the transactions between the sender and the receiver taking into consideration the gap between each couple of transactions and the amount of each. The score calculation is based on a decreasing function of power two, and a factor τ. The idea is that a transaction looses its importance over time. After a given time τ, the score of the transaction is half its initial value. For a couple of nodes, the equation of this score is given by:

$$w(v_i, v_j) = \sum_{i=1}^{\#(TA)} transaction\,amount(v_i, v_j) \ * 2^{\frac{-(t-t_i)}{\tau}} \quad (5)$$

For a couple of nodes, this sum is calculated on the basis of all the transactions exchanged with an amount $transaction\,amount_{v_i,v_j}$, where t is the date of each transaction, t_f is the date of the last exchanged transaction and τ equals to 30 days.

V. Experimental protocol

A. Methodology

The main purpose of this work is to investigate the impact of simplifying a multigraph into a weighted graph, in the context of overlapping communities detection. We aim to identify the best method of simplification while testing different weighting functions. The assessment of this impact will be achieved through the application of community detection methods on both, the original multigraph, and the different generated graphs, and through the examination of the structural properties of the communities formed.

Our methodology is as follows: starting from the initial multigraph, five weighted graphs will be created, using the five weighting functions described in section IV. We will also create a simple graph, where edges have equal weights of 1. This graph will serve as a reference to test if weights are as important as we anticipate. Three overlapping community detection algorithms will be applied to each of these graphs. The implemented algorithms are Aslpaw [19], Lswl-plus [20], and wCommunity [21].

For our tests, the crucial factor for choosing an algorithm is its ability to be applied on multigraphs, and the fact that it takes into consideration edges directions ans weights.

The community detection algorithms will be applied to the original multigraphs, as well as the simple graph and the weighted graphs. The overlapping communities generated will be then compared to the ground-truth communities, through suitable evaluation metrics. Results will be subsequently compared, in order to study the impact of transforming a multigraph to a simple graph in view of community detection. We also will be able to validate the effectiveness of each simplification method.

B. Evaluation metrics

One of the biggest difficulties related to community detection in social networks, is the ability to evaluate the obtained results. Evaluation is a real issue for real networks, where only little data are provided. In the last years, a lot of work was performed regarding evaluating metrics for community detection [22], [23]. Despite the large number of community detection evaluation metrics in the literature, most of the work is focused on disjoint communities. But an increasing attention has turned to overlapping communities' detection evaluation over the last years [23], [24]. An adaptation of the most known metrics to overlapping applications were introduced, even if their efficiency is yet debatable. These metrics can formally be classified into two categories: intrinsic and extrinsic metrics.

Intrinsic metrics evaluate structural properties of the identified communities. They assess how similar the elements of each community are, and how they differ from elements in other communities, given a specific metric. The most common intrinsic quality metrics available for overlapping communities are modularity [25], conductance [26], and the average internal degree:

- Modularity. It is a chance corrected metric that measures the difference between the number of within-communities edges in the network, and the expected value of the same quantity in a random network [25]. Modularity was introduced in order to assess the relevance of a given community structure compared to a random structure. A value over 0 indicates a deviation from randomness, and values close to 1 indicates a strong community structure. The most popular algorithms of community detection such as *Girvan-Newman* [27] and *Louvain* [28] are based on modularity maximization.
- Conductance. The conductance of a community is the ratio between the number of edges that point outside the community and the minimum between the number of edges with an endpoint in the community or the number of edges that do not have an endpoint in the community. Yang and Leskovec [29] have shown that conductance is a good metric for evaluating communities of real world graphs. The conductance value varies between 0 and 1. A lower conductance indicates that the community is more "well-knit". The conductance of the graph is the average of the conductance of each community.
- Average internal degree: Average degree is the average number of edges per node in the graph. It is given by the division of

the total number of edges by the number of nodes. The average internal degree is the average degree within the community. Hence, the internal average degree of communities' partition is given by the average of all the internal average degrees of individual communities.

Extrinsic metrics on the other hand evaluate how the resulted communities are comparable to a *gold standard* also called *ground-truth*. These metrics are also called information recovery metrics because they measure the ability of algorithms to recover information from the ground-truth. For real-world networks this ground-truth is not always available. In such case, we consider that evaluation of the performance of an overlapping community detection algorithm is less straightforward. Therefore, in this work we only consider synthetic networks where the ground-truth is available. Thus, information recovery metrics compare two sets of communities (not necessarily the same number) based on different criteria. The most popular information recovery metrics are the Normalized Mutual Information (NMI) [30], and the F1-score [10].

In order to overcome the shortcomings of existing metrics, we proposed four information recovery metrics in a previous submitted work, for comparing communities that overlap. More specifically, we presented metrics that aim to compare the similarity of an overlapping community detection result, to a given ground-truth. These metrics are the *inclusion rate*, the *coverage rate*, the *overlapping rate*, and the *distribution rate*, which will be detailed just below. All the proposed metrics vary between minimum 0 and maximum 1. The used extrinsic metrics are:

- ONMI: The Overlapping Normalized Mutual Information (ONMI) is an adjustment of the Normalized Mutual Information (NMI) for overlapping communities. The NMI metric has become one of the most popular metrics when it comes to evaluate the relevance of communities [30]. Based on information theory, the NMI measures the similarity between two partitions. NMI value varies between 0 and 1. Lancichinetti *et al.* [31] proposed a variation of normalized mutual information fitted to overlapping communities, namely ONMI.
- Average F1-score: The average F1 score is the mean of the F1-scores of the best matching ground-truth community to each detected community and the F1-scores of the best-matching detected community to each ground-truth community [10]. The $F1$ score is given by the harmonic mean of Precision and Recall.
- Inclusion rate: The inclusion rate is a metric that is meant to measure the embeddedness of result communities into ground-truth communities. The basic idea behind this metric was the need of a measure that estimates the amount of representativeness of the result communities compared to the ground-truth ones. For a given result community R_i, the individual inclusion rate is given by the maximum precision rate:

$$\begin{aligned} Inclusion\,rate(R_i) &= \max_j(precision(R_i, G_j)) \\ &= \max_j\left(\frac{R_i \bigcap G_j}{|R_i|}\right) \end{aligned} \tag{6}$$

The overall inclusion rate is then defined by the ratio of a weighted sum of the individual inclusion rates divided by the sum of the resulting communities sizes. The weight we chose for the sum is the size of individual result communities:

- Coverage rate. While the inclusion rate regards similarity from the result communities perspective, the coverage rate considers it from the ground-truth angle. Our purpose was to identify two complementary metrics, that account for analogous similarity perspectives. As the inclusion rate is based on the maximum of the precision, the coverage rate is a function of the maximum recall. For a given ground-truth community G_j, the individual coverage rate is given by the maximum recall rate:

$$\begin{aligned} Coverage\,rate(G_j) &= \max_i(recall(R_i, G_j)) \\ &= \max_i\left(\frac{R_i \bigcap G_j}{|G_j|}\right) \end{aligned} \tag{7}$$

The overall coverage rate is then defined by the ration of a weighted sum of the individual coverage rates by the sum of the ground-truth communities sizes. We chose the size of ground-truth communities to weight the sum:

- Overlapping rate. The overlapping rate is a metric that calculates the total number of overlapping nodes between communities. For every pair of communities, the overlapping rate is calculated by dividing the number of common nodes they share by the smallest size between the two communities. The overlapping rate for a couple of result communities for example is given by:

$$Overlapping\,rate(R_i, R_j) = \frac{|C_i \bigcap C_j|}{min(|C_i|, |C_j|)} \tag{8}$$

For a given partition, the overlapping rate is given by the mean of the overlapping rates of the pairs of communities it contains. Since our metric is an information recovery metric, the overlapping rate between two partitions R and G, is given by the equation 9 below:

$$\begin{aligned} &Overlapping\,rate(R, G) = \\ &1 - |Overlapping\,rate(R) - Overlapping\,rate(G)| \end{aligned} \tag{9}$$

- Distribution rate. The distribution rate compares the number of communities each node belongs to in the result communities and in the ground-truth. For each node i, the distribution rate is given by:

$$Distribution\,rate_i = |n_g(i) - n_r(i)| \tag{10}$$

Where n_g and n_r are the number of communities in the ground-truth and the result to which the node i belongs. The distribution rate is given by:

$$Distribution\,rate = \exp\left(-\frac{\sum_i Distribution\,rate_i}{|V|}\right) \tag{11}$$

- Average score: Each one of our proposed metric considers a structural aspect of the communities acquired by the community detection algorithm. Accordingly, we impose that a good set of result should have good scores for all the four metrics. All four metrics must be considered concurrently to fully understand the results, as some metrics may give good scores on some bad results. Therefore, an average score of the inclusion rate, the coverage rate, the overlapping rate and the distribution rate should be computed.

C. Data description

In the scope of this study, we are provided with a high volume of data. However, due to the sensitive nature of financial data, and the strict organisation internal policies to protect users confidentiality, we cannot disclose any information about users. Therefore simulated data will be used for our experiments. Simulated data are self-sufficient data aspiring to have comparable statistical properties as real one. Simulation is also useful to assess the outcome of tested algorithms, before applying them on real data.

Real data may be random and the lack of a ground-truth leads to a limited understanding of community structure in such networks. Therefore, testing community detection algorithms on real-world networks may not be conclusive due to the lack of the ground-truth. The generated network contains $10,000$ nodes and more than $350,000$ transactions over a one-year span. The ground truth includes $3,626$ communities having an average size of 20 nodes.

VI. Experimentation

A. Results

In order to compare different reduction methods, as previously mentioned, we apply three overlapping community detection algorithms (Aslpaw, Lswl-plus, and wCommunity) on the synthetic data.

The results of extrinsic and intrinsic metrics applied on the communities generated are showed in tables I to III. In these tables are displayed metrics' results for the multigraph and for the five reduced graphs.

For Aslpaw (table I) the extrinsic metrics show that the community detected on the multigraph yields good results when compared to the ground truth. The inclusion and coverage rates are both high. This indicates that communities found by the algorithm are an acceptable match of the ground truth. On the other hand, the results of our proposed metrics on the simple graph present the best the results, when compared to the multigraph, and to the other reduced graphs. The simple graph has the best NMI score. The multigraph has the best f1-score.

For Lswl-plus table (II), results show disparate behaviors on the multigraph compared to the reduced graphs. For multigraph, the inclusion and coverage rates are both high which imply an acceptable community structure discovered relative to the ground truth. For the reduced graphs, we observe a low inclusion rates and very high coverage rates. This suggests that discovered communities present a case of under-segmentation. This occurs when the algorithm generates communities that combine multiple ground-truth communities into one extensive community. The distribution rate has the best value with the simple graph. The overlapping rates are very close, but the best value is met with the multigraph. The multigraph has also the best average score. The simple graph has the highest NMI result, and the multigraph has the best f1-score result.

For the last algorithm wCommunity, results, displayed in (III), show a comparable behavior to Aslpaw. The inclusion and coverage rates are both relatively high. The simple graph has the best average score, but with close results to the occurrence and the mean graphs. The multigraph has the best NMI and f1-score values.

From a structural perspective, if we consider the conductance and the overlapping modularity results, we observe that the reduced graphs score generally better than the multigraph.

B. Discussion

In the previous section, we presented the results of intrinsic and extrinsic evaluation metrics applied on result communities of three different overlapping communities detection algorithms. The goal of our experiments is twofold. The first aim is to determine if a multigraph reduction allows to reach better results regarding the detection of overlapping communities. Since we consider different reduction approaches based on different calculations of edges weights, the second goal of the experiments is to identify which reduction method achieves better overlapping community detection results. The first observation that can be drawn from the results is that the algorithm Lswl-plus produces better scores for the multigraph compared to the other reduced graphs. This may be an indication that this community detection algorithm is more efficient on multigraphs than on weighted graphs and has been specifically designed for multigraphs and not weighted graphs. For extrinsic evaluation metrics, while the multigraph has the bigger inclusion rate, the reduced graphs have for the major part a better coverage rate which means that the ground truth communities are better represented in the result communities.

One of the reason that led us to propose new evaluation metrics, is the lack of coherence and the difficulty of interpretation wen it comes to standard metrics such as NMI and f1-score. Communities with quite different structures and distribution, can have close or the same NMI measures. Consequently, it is necessary to analyze the topological properties of community structure.

As for intrinsic quality metrics, results highlight a structural disparity between communities generated from the multigraph, as opposed to the communities generated from weighted graphs. The values of the average internal degrees indicate that the multigraph communities are bigger than the rest of the communities. However, the overlapping modularity values (and some conductance values) are overall superior for the reduced graphs, which underlines a stronger community structure.

Based on the experimental results, the first conclusion that can be drawn is that the choice of the algorithm has

TABLE I
EXTRINSIC AND INTRINSIC METRICS FOR *Aslpaw*

	Aslpaw						
	Multigraph	Simple	Occurrence	Sum	Mean	Monthly-mean	temporal-score
Inclusion rate	**0.837**	0.706	0.789	0.785	0.748	0.773	0.783
Coverage rate	0.506	**0.675**	0.535	0.519	0.606	0.551	0.447
Distribution rate	0.074	0.073	0.077	0.077	0.075	0.078	**0.082**
Overlapping rate	0.99	0.99	0.99	0.99	0.99	0.99	0.99
Average score	0.602	**0.611**	0.598	0.593	0.605	0.598	0.576
ONMI	0.230	**0.278**	0.188	0.172	0.222	0.190	0.125
F1-score	**0.266**	0.235	0.246	0.245	0.245	0.244	0.233
Average internal degree	29.16	3.65	4.01	3.81	4.79	3.99	2.99
Conductance	**0.321**	0.460	0.521	0.54	0.42	0.52	0.62
Overlapping modularity	0.0109	**0.0952**	0.0407	0.0318	0.0504	0.0381	0.0187

TABLE II
EXTRINSIC AND INTRINSIC METRICS FOR *Lswl-plus*

	$Lswl - plus$						
	Multigraph	Simple	Occurrence	Sum	Mean	Monthly-mean	temporal-score
Inclusion rate	**0.719**	0.202	0.194	0.187	0.183	0.194	0.223
Coverage rate	0.521	**0.825**	0.816	0.818	0.816	0.814	0.813
Distribution rate	0.0678	**0.107**	0.0166	0.012	0.00891	0.017	0.0316
Overlapping rate	**0.998**	0.997	0.992	0.992	0.991	0.993	0.993
Average score	**0.576**	0.533	0.505	0.502	0.500	0.505	0.515
ONMI	0.151	**0.217**	0.153	0.168	0.155	0.157	0.159
F1-score	**0.23**	0.077	0.068	0.066	0.063	0.068	0.077
Average internal degree	23.21	5.17	9.35	9.38	9.49	9.15	9.23
Conductance	0.39	0.23	**0.15**	**0.15**	**0.15**	0.16	0.16
Overlapping modularity	0.069	**0.14**	0.057	0.058	0.0517	0.055	0.060

an important influence on the comparison of overlapping communities detection on a multigraph versus a weighted graph. For this reason, one of our future works as part of this PhD will be based on the development of an adequate overlapping community detection algorithm centered on the edges' weights.

While the reduction of the multigraphs may allow to simplify the calculations (dividing the number of edges by five) and to implement community detection algorithms that do not operate on multigraphs, it doesn't necessarily lead to the best result due to information loss.

The comparison of the different reduction methods and different edge weights computations suggest that their performances are quite comparable. Thereafter, the usage of one of these reduction methods would depend on the goal of the study and the type of data explored.

Finally, it can be asserted that in order to ensure better results, we need to consider the overlapping community detection algorithm, the reduction method, and the evaluation metrics concurrently.

VII. CONCLUSION

In this paper, we addressed the question of multigraph reduction into weighted graph. A multigraph is a special type of graph where nodes could have multiple edges connecting them. The multigraph studied on this paper is an example of synthetic transactional network where nodes represent customer, and edges represent the transactions' connecting them. The comparison based on intrinsic and extrinsic metrics on the result of overlapping community detection applied on the multigraph as well as on six weighted graphs show that communities of weighted graphs are on general more similar to the ground truth and are structurally better than those of the multigraph. Nevertheless, we highlight that the choice of the weighting function should rely on the context and the purpose of the study.

REFERENCES

[1] L. Peel, D. B. Larremore, and A. Clauset, "The ground truth about metadata and community detection in networks," *Science Advances*, vol. 3, no. 5, p. e1602548, 2017.

TABLE III
EXTRINSIC AND INTRINSIC METRICS FOR *Wcommunity*

	Wcommunity						
	Multigraph	Simple	Occurrence	Sum	Mean	Monthly-mean	temporal-score
Inclusion rate	**0.815**	0.803	0.804	0.791	0.780	0.776	0.737
Coverage rate	0.576	0.597	0.585	0.577	**0.613**	0.602	0.592
Distribution rate	0.076	0.072	0.079	0.081	0.075	0.077	**0.084**
Overlapping rate	0.998	0.998	0.999	0.999	0.998	0.998	0.999
Average score	0.616	**0.618**	0.617	0.612	0.617	0.613	0.603
ONMI	**0.250**	0.202	0.213	0.215	0.223	0.212	0.213
F1-score	**0.260**	0.245	0.249	0.244	0.244	0.245	0.232
Average internal degree	27.6	1.47	3.32	3.51	3.74	3.72	4.62
Conductance	**0.44**	0.78	0.65	0.63	0.62	0.61	0.51
Overlapping modularity	0.0165	0.048	0.049	0.049	0.058	0.053	**0.079**

[2] G. Rossetti, L. Pappalardo, and S. Rinzivillo, *A Novel Approach to Evaluate Community Detection Algorithms on Ground Truth*. Springer International Publishing, 2016.

[3] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3, pp. 75–174, 2010.

[4] T. Shafie, "A multigraph approach to social network analysis," *journal of Social Structure*, vol. 16, pp. 1–22, 2015.

[5] S. Centellegher, G. Miritello, D. Villatoro, D. Parameshwar, B. Lepri, and N. Oliver, "Mobile Money: Understanding and Predicting its Adoption and Use in a Developing Economy," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 4, p. 157, 2018.

[6] R. Di Clemente, M. Luengo-Oroz, M. Travizano, S. Xu, B. Vaitla, and M. C. González, "Sequences of purchases in credit card data reveal lifestyles in urban populations," *Nature Communications*, vol. 9, no. 1, 2018.

[7] S. E. Ayeb, B. Hemery, F. Jeanne, and E. Cherrier, "Community detection for mobile money fraud detection," in *Seventh International Conference on Social Networks Analysis, Management and Security (SNAMS)*, 2020, pp. 1–6.

[8] P. Kim and S. Kim, "A detection of overlapping community in mobile social network," in *Proceedings of the 29th Annual ACM Symposium on Applied Computing*, 2014, pp. 175–179.

[9] A. Ponomarenko, L. Pitsoulis, and M. Shamshetdinov, "Overlapping community detection in networks based on link partitioning and partitioning around medoids," *PLOS ONE*, vol. 16, no. 8, p. e0255717, 2021.

[10] J. Yang and J. Leskovec, "Overlapping community detection at scale: a nonnegative matrix factorization approach," in *ACM International conference on Web search and data mining*, 2013, pp. 587–596.

[11] J. Xie, S. Kelley, and B. K. Szymanski, "Overlapping community detection in networks: The state-of-the-art and comparative study," *ACM Computing Surveys (csur)*, vol. 45, no. 4, pp. 1–35, 2013.

[12] S. Y. Bhat and M. Abulaish, "Overlapping social network communities and viral marketing," in *International Symposium on Computational and Business Intelligence*, 2013, pp. 243–246.

[13] Y. Li and S. Zhang, "A community discovery and TrustRank based approach for spammer ranking," in *International Conference on Culture-oriented Science Technology (ICCST)*, 2020, pp. 279–283.

[14] E. E. Papalexakis, L. Akoglu, and D. Ience, "Do more views of a graph help? community detection and clustering in multi-graphs," in *Proceedings of the 16th International Conference on Information Fusion*, 2013, pp. 899–905.

[15] M. Gjoka, C. T. Butts, M. Kurant, and A. Markopoulou, "Multigraph sampling of online social networks," *IEEE Journal on Selected Areas in Communications*, vol. 29, no. 9, pp. 1893–1905, 2011.

[16] G. Cormode and S. Muthukrishnan, "Space efficient mining of multigraph streams," in *Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*. Association for Computing Machinery, 2005, pp. 271–282.

[17] D. Duan, Y. Li, Y. Jin, and Z. Lu, "Community mining on dynamic weighted directed graphs," in *Proceedings of the 1st ACM international workshop on Complex networks meet information & knowledge management*, 2009, pp. 11–18.

[18] J. Majmudar and S. Vavasis, "Provable overlapping community detection in weighted graphs," *Advances in Neural Information Processing Systems*, vol. 33, pp. 19 028–19 038, 2020.

[19] J. Xie, B. K. Szymanski, and X. Liu, "SLPA: Uncovering overlapping communities in social networks via a speaker-listener interaction dynamic process," in *2011 IEEE 11th International Conference on Data Mining Workshops*, 2011, pp. 344–349.

[20] M. Zafarmand, "Fast local community discovery: Relying on the strength of links." https://github.com/charlespwd/project-title, 2021.

[21] D. Chen, M. Shang, Z. Lv, and Y. Fu, "Detecting overlapping communities of weighted networks via a local algorithm," *Physica A: Statistical Mechanics and its Applications*, vol. 389, no. 19, pp. 4177–4187, 2010.

[22] T. Chakraborty, A. Dalmia, A. Mukherjee, and N. Ganguly, "Metrics for community analysis: A survey," *ACM Computing Surveys*, vol. 50, no. 4, pp. 1–37, 2017.

[23] M. K. Goldberg, M. Hayvanovych, and M. Magdon-Ismail, "Measuring similarity between sets of overlapping clusters," in *IEEE Second International Conference on Social Computing (SocialCom)*, 2010, pp. 303–308.

[24] A. Lutov, M. Khayati, and P. Cudré-Mauroux, "Accuracy evaluation of overlapping and multi-resolution clustering algorithms on large datasets," in *IEEE International Conference on Big Data and Smart Computing (BigComp)*, 2019, pp. 1–8.

[25] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Physical Review E*, vol. 69, no. 2, 2004.

[26] H. Almeida, D. Guedes, W. Meira, and M. J. Zaki, "Is there a best quality metric for graph clusters?" *Machine Learning and Knowledge Discovery in Databases*, pp. 44–59, 2011.

[27] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, 2004.

[28] 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, 2008.

[29] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," *Knowledge and Information Systems*, vol. 42, no. 1, pp. 181–213, 2015.

[30] L. Danon, A. Díaz-Guilera, J. Duch, and A. Arenas, "Comparing community structure identification," *Journal of statistical mechanics: Theory and experiment*, vol. 2005, no. 9, p. P09008, 2005.

[31] A. Lancichinetti, S. Fortunato, and J. Kertesz, "Detecting the overlapping and hierarchical community structure of complex networks," *New journal of Physics*, vol. 11, no. 3, 2009.