Towards Preventing Neighborhood Attacks: Proposal of a New Anonymization's Approach for Social Networks Data

Requi Djomo[1(✉)] and Thomas Djotio Ndie[2]

[1] National Polytechnic School of Douala (ENSPD), University of Douala, Douala, Cameroon
[2] National Polytechnic School of Yaounde (ENSPY), University of Yaounde, Yaounde, Cameroon

Abstract. Anonymization is a crucial process to ensure that published social network data does not reveal sensitive user information. Several anonymization approaches for databases have been adopted to anonymize social network data and prevent the various possible attacks on these networks. In this paper, we will identify an important type of attack on privacy in social networks: "neighborhood attacks". But it is observed that the existing anonymization methods can cause significant errors in certain tasks of analysis of structural properties such as the distance between certain pairs of nodes, the average distance measure "APL", the diameter, the radius, etc. This paper aims at proposing a new approach of anonymization for preventing attacks from neighbors while preserving as much as possible the social distance on which other structural properties are based, notably APL. The approach is based on the principle of adding links to have isomorphic neighborhoods, protect published data from neighborhood attacks and preserve utility on the anonymized social graph. Our various experimental results on real and synthetic data show that the algorithm that combines the addition of false nodes with the addition of links, allows to obtain better results compared to the one based only on the addition of links. They also indicate that our algorithm preserves average distances from the existing algorithm because we add edges between the closest nodes.

Keywords: Anonymization · Social network · Neighborhood attacks · Confidentiality · Graph isomorphism · APL

1 Introduction

Currently, more and more social network data is made available to the public in one way or another, protecting personal information while publishing social network data is becoming a very important concern. With some local knowledge about the individuals in a social network, an adversary can easily invade the privacy of some victims. Is it possible that posting social media data, even with anonymous individuals, still threatens privacy? Many social networks collect confidential information about their users, information

Z. Deze et al. (Eds.): BDTA 2020/WiCON 2020, LNICST 371, pp. 195–208, 2021.
https://doi.org/10.1007/978-3-030-72802-1_14

that could potentially be misused. The development of online social networks and the publication of social network data has created a risk of personal information leaking from individuals (example: salary, illness, connection to a specific group, etc.) (See Fig. 1. This requires the preservation of privacy before such data is published. In this paper, we focus on an important type of attack on privacy in social networks: "neighborhood attacks". If an adversary has some knowledge about a target victim's neighbors and the relationships between those neighbors, the victim can be re-identified even if the victim's identity is preserved using conventional anonymization techniques.

Most of the previous privacy studies can only deal with relational data, and cannot be applied directly to social media data. However, Bin Zhou and Jian Pei [1] have proposed a method of anonymizing a social network to prevent re-identification of nodes through structural "neighborhood" information, which is an initiative in this direction and which provides a practical solution to the problem. Subsequently, other research works were also proposed to solve the same problem as in [2–4].

The in-depth study of the neighborhood attack gave us a good understanding of the anonymization approach proposed by Zhou and Pei against these attacks. This approach, which works based on the principle of adding links to have isomorphic neighborhoods, considerably preserves privacy against neighborhood attacks but can significantly modify the structural properties of the original graph and therefore its potential utility. Our issue falls within this context, namely the protection of privacy while preserving a very important structural property, namely "APL (Average Shortest Path Length)". Preserving such a property in anonymization could be extremely essential thereafter for the analysis of graphs of anonymized networks.

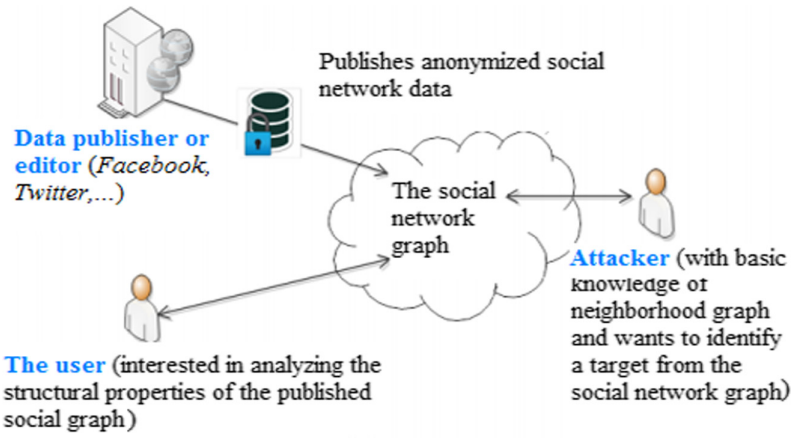

Fig. 1. Publication of social network data

The rest of this work is organized as follows: we start in Sect. 2 with an illustration of the neighborhood attack to show how an attacker can identify a person in the published graph using their neighborhood knowledge. Subsequently, we discuss the limits of the existing solution to prevent this attack. We present our contribution to improve this solution in Sect. 3. Our approach to social network anonymization that preserves as much as possible the ownership of the mean distance or APL is presented in Sect. 5. Finally, Sect. 7 is devoted to the conclusion, and then some tracks for future work are discussed.

2 Illustration of Neighborhood Attacks

As a concrete example, let us take a synthesized social network of "friends" shown in Fig. 2 (a). Each vertex in the network represents a person. The sensitive attribute associated with each node in the network represents a disease, and the non-sensitive attribute represents each person's occupation. An edge connects two people who are friends. Let us suppose that the network is to be published. To preserve privacy, is it sufficient to delete all identities as shown in Fig. 2 (b) (i.e. perform naive anonymization [11], where identifiers are replaced with random numbers)?

Unfortunately, if an adversary has some knowledge of an individual's neighbors, privacy may be disclosed. If an opponent wants to find information about "Walid" and knows that he has two friends who know each other and, two other friends who do not know each other, i.e. it knows the neighborhood graph of "Walid" as shown in Fig. 2 (c), then the vertex "2" representing "Walid" can be uniquely identified in the network since no other vertex has the same graph of neighborhood. The adversary can thus know that "Walid" suffers from epilepsy. This represents an intrusion into the privacy of "Walid". Likewise, "Lyes" can be identified in Fig. 2 (b) if the adversary knows the neighborhood graph of "Lyes".

Identifying Individuals in Published Social Networks Violates Privacy.

In this example, by identifying "Walid" and "Lyes", an adversary may even know from the published network (Fig. 2 (b)) that "Walid" and "Lyes" share a friend in common.

Now, let us assume the opponent wants to find information about "Lina" and knows that she has two friends who do not know each other in the network. Using this knowledge, he tries to find it in the network. There are 4 vertices in the network which have the same neighborhood: 5, 7, 3 and 4 as shown in Fig. 2 (b). "Lina" can be any of these. Thus, "Lina" cannot be identified in the social network with a probability greater than 1/4. If each node in the social network cannot be identified with a probability greater than 1/k, the network is said to follow the principle of "k-anonymity" [5].

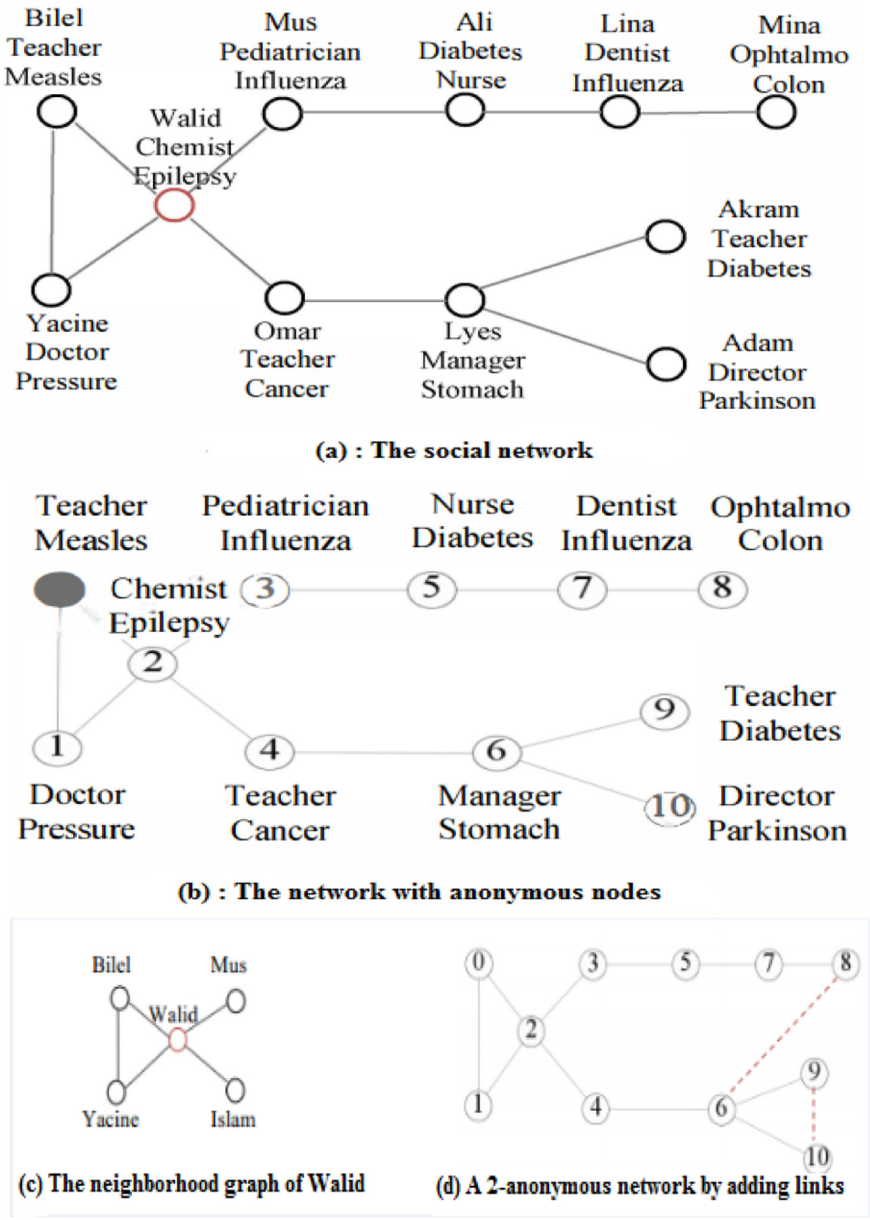

Fig. 2. Neighborhood attacks in a social network [12]

3 Contribution

To protect privacy satisfactorily, one solution is to ensure that any individual cannot be identified correctly in the anonymized social network with a probability greater than

1/k, where k is a user-specified parameter carrying the same characteristics of the k-anonymity model of L. Sweeney [5].

In the example of Fig. 2, an anonymous "2-neighborhood" graph of Fig. 2 (a) generated by adding links can be published. By adding two false links, one connecting "Lyes" and "Mina" and the other connecting "Akram" and "Adam", the neighborhood graph of each vertex in Fig. 2 (d) is no longer unique. An adversary with the knowledge of the neighborhood of a node, always gets at least two candidate nodes, so he cannot identify an individual in this anonymous graph with a probability greater than 1/2.

Zhou and Pei [1] took the initiative to address the problem of preserving privacy in the publication of social networks against neighborhood attacks, and proposed an anonymization method based on the addition of links, and subsequently, other research works [2–4] have appeared which are based on the solution of Zhou and Pei [1].

For example by connecting "Lyes" and "Mina" the distance between nodes 6 and 8 is changed from 6 to 1 in Fig. 2 (d). We note that this addition of link significantly modified the value of the distance and therefore any analysis (data mining) on this data could obtain erroneous or invalid conclusions. The advantage of the method of adding links already proposed keeps the nodes in the original graph unchanged, however, it can greatly affect the structure of the graph. This method can sometimes modify the distance properties appreciably for example by connecting two distant nodes which belong to two different communities.

To better explain this example, consider that the structure of communities can be detected from the relationships of friends in the social network. Suppose that 6 and 8 are members of two different communities in the original graph, and the communities are far from each other. By connecting 6 to 8, these communities can become very close or merge to form a single community.

So relying solely on adding links may not be a good solution to keep data useful. To solve this problem, we propose to preserve important properties of graphs, such as distance, the addition of false nodes in the graph. For example, if we simply add a false node to the graph in Fig. 2(a), we can also generate an anonymous 2-neighborhood graph as shown in Fig. 3. In this figure, the distances between the nodes of the original graph haven't changed much.

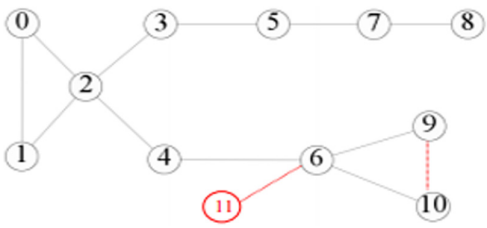

Fig. 3. An anonymous 2-neighborhood graph by adding false nodes

4 Some Concepts of Social Graphs

Before presenting our approach to anonymization, it is necessary to define some basic concepts used in our work.

4.1 Neighborhood and d-Neighborhood of a Vertex

In a social graph G, the neighborhood of a vertex u ∈ V (G) is the subgraph induced by the neighbors of u, denoted $Neighbor_G(u) = G\ (V_u)$ where $V_u = \{v \mid (u, v) \in E\ (G)\}$ [1]. The neighborhood graph of a vertex u includes all vertices that are in the distance "d" from the vertex u [2].

4.2 Neighborhood Component

In a social network G, a subgraph C of G is a neighborhood component of u ∈ V (G) if C is a maximal connected subgraph in $Neighbor_G(u)$. Figure 4 shows $Neighbor_G(u)$, the neighborhood of a vertex u, which contains three neighborhood components C_1, C_2 and C_3. Clearly, the neighborhood of a vertex can be divided into neighborhood components.

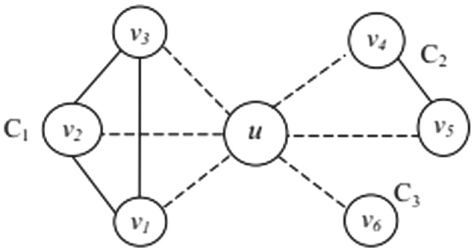

Fig. 4. Neighborhood and neighborhood components (the dotted edges are just for illustration and are not in the neighborhood subgraph) [1]

The following table summarizes the notations used in this work (Table 1):

Table 1. Notations used.

Symbol	Description
G	The initial graph modeling a social network
V(G), E(G)	V: The set of vertices in the graph G, E: the set of edges
G′	The anonymized graph of the social graph G
$Neighbor_G(u)$	Neighborhood of vertex u
$C_i(u)$	The component number i in the neighborhood of the vertex u
$\lvert V(C_i)\rvert$, $\lvert E(C_i)\rvert$	The number of vertices and the number of edges in the component C_i

4.3 Graph Isomorphism

Let be two graphs: G_1 (V_1, E_1) and G_2 (V_2, E_2), where $|V_1| = |V_2|$, G_1 is isomorphic to G_2 if and only if there exists at least one bijective function between V_1 and V_2: f: $V_1 \rightarrow V_2$, such that $\forall(u, v) \in E_1$, there exists an edge ((f (u), f (v)) $\in E_2$. For example, the two graphs below (a) and (b) in Fig. 5 are isomorphic [6].

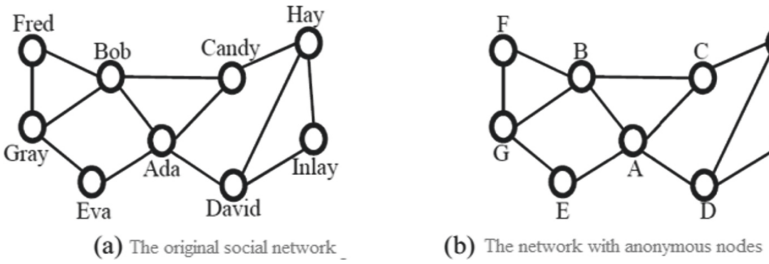

Fig. 5. Graph isomorphism

4.4 The Subgraph Isomorphism

For two graphs G1 = (V_1, E_1) and G_2 = (V_2, E_2), G_1 and G_2 are isomorphic subgraph if G_1 contains an isomorphic subgraph to G_2. As shown by the graphs from (a) to (c) in Fig. 6, they can find the corresponding isomorphic subgraphs in Fig. 5 (b) [6].

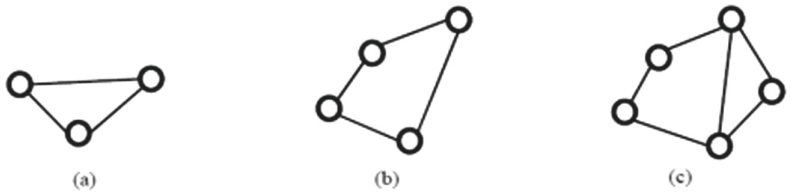

Fig. 6. Example of subgraph isomorphism [6]

4.5 Usefulness of the Anonymized Graph and Structural Properties

When the social network graph is changed, it is a big challenge to balance between preserving privacy and losing the usefulness of the data. In the context of our study, we are interested in utility measures reflecting the structural properties of a social graph G = (V, E), in particular the APL. These characteristics are generally used by analysts, for example to study influence, and forms of communication in social networks, to do viral marketing or to study the patterns of the spread of information and disease, etc. [7]. The result obtained was, on average, six intermediate nodes are sufficient to send a letter to a target individual [8].

5 Proposal For A New Approach to Anonymization

A major challenge in anonymizing a social network is that changing the labels of vertices and adding edges as well as false vertices can affect the neighborhoods of other vertices as well as the properties of the network. It is well recognized that the following two properties are often retained in social networks. These properties help us in the design of our method of anonymization.

Property 1: "the distribution of the degrees of the vertices follows a power law": the distribution of the degrees of the vertices in a large social network often follows the power law [9]. These degree distributions were identified in various social networks, including the Internet, biological networks and co-author networks.

Property 2: "the small-world phenomenon" [8]. It is also popularly known as "six degrees of separation", which states that large social networks in practice often have surprisingly small average diameters.

Our social network anonymization method treats vertices in descending order of degree, and uses the above two properties.

In this paper, the idea is to develop a new technique of social graph anonymization that is not only based on adding links but also on adding false nodes in order to reach a compromise between the preservation of privacy and the resulting utility of the anonymized graph.

Our major contribution is to propose a new efficient social network anonymization algorithm called "AnonSN", which not only preserves the privacy of individuals present in the published social graph, preventing an attacker from being able to identify a user using the neighborhood knowledge, but also which maintains as much as possible the different structural properties of the original graph, including distance. This algorithm is based on the combination of the principle of "adding links" and that of "adding false nodes".

The proposed anonymization approach ensures that for any modification made to the original graph, the property of mean distance between the vertices involved is as close as possible to that in the original graph. We have also proposed a new formula for calculating the cost of anonymization. In summary, we have made the following main contributions:

1. Proposal of a new formula for calculating the cost of anonymization to determine the neighborhoods of the vertices that can be anonymized together.
2. Proposal of a new approach to anonymization based on the addition of false nodes in addition to the addition of links.
3. Comparison of the performance of the proposed approach of adding false nodes to Zhou and Pei's approach according to the preservation of the APL.

The architecture of our anonymization approach is illustrated in Fig. 7 below:

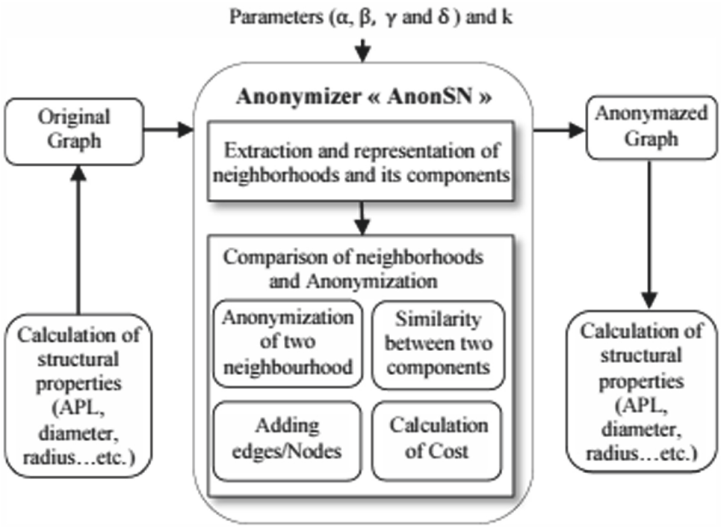

Fig. 7. Architecture of the anonymization approach

5.1 Extraction and Representation of Neighborhoods and Neighborhood Components

The neighborhoods of all the vertices in the graph G are extracted and the different components are separated. To facilitate comparisons between neighborhoods of different vertices, including isomorphism tests that will be frequently performed in anonymization, we have chosen adjacency matrices to represent the different neighborhood components, such as:

$$M[i,j] = \begin{cases} 1, \ if \ \ (i,j) \in E \\ 0, \ otherwise \end{cases} \tag{1}$$

Using the "Walid" neighborhood example in Fig. 2(c), the "Walid" neighborhood components and the adjacency matrices that represent them are shown in Fig. 8 below:

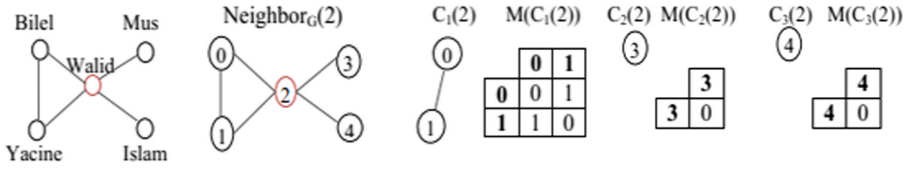

Fig. 8. Neighborhood and neighborhood components of "Walid"

5.2 Measuring the Quality of Anonymization "Cost of Anonymization"

The main goal of social network anonymization algorithms is to develop efficient heuristics to ensure a balance between preserving the structure of the original graph and the

privacy of individuals. The strength of an anonymization algorithm can be measured in terms of the loss of information.

In our anonymization model, we have proposed a new formula for calculating the cost of anonymization. There are three ways to anonymize vertex neighborhoods: "generalize vertex labels", "add edges" and "add false nodes". Each of the three methods leads to some loss of information.

Using the social graph presented in Fig. 2 we illustrate on the following table, the values of some structural properties of the graph anonymized using the two anonymization methods: Zhou and Pei edge addition method and our proposed method (Table 2):

Table 2. Structural properties by adding vs links by adding false vertices.

	Original graph	Zhou and Pei	Proposed method
Diameter	7	4	7
Radius	4	3	4
APL	3.109	2.4727	3.1818
Density	0.200	0.236	0.197

We notice that the values of the structural properties of the graph anonymized with the proposed method of adding false nodes are closer to the values of the properties of the original graph when compared with those of the method of adding links, and thus they are better preserved.

6 Experiments

In this part, we describe the different experiments performed on real and synthetic data sets. Thus, we discuss the results to assess and illustrate the performance of our proposed approach by comparing our approach to the existing approach of Zhou and Pei [1].

6.1 Result of Calculation of the Structural Properties: (Synthetic and Real Graphs)

We calculate the values of the structural properties for the original graph and the resulting anonymized graphs. To calculate the values of all the structural properties, we use the Gephi software [10]. Finally, during the last step of our experiment, we compare the values of the structural properties measured for the original social graph with those obtained for the anonymized graphs. The results obtained are shown in Tables 3 and 4. On each table, the row represents the values of a structural property in the original graph and the one anonymized using the two algorithms. The context is defined by the couple: (number of vertices, number of edges).

Table 3. Structural properties of the anonymized graphs generated for the synthetic graph 40 nodes.

	Original graph	Zhou et Pei	"AnonSN"
Context	(40,51)	(40,56)	(41,54)
Nber added edges	/	5	3
Number of false vertices added	/	/	1
APL	5.347435	4.485897	5.332926
Diameter	13	12	13
Radius	7	7	7
Medium degree	2.550	2.800	2.634
Density	0.065	0.072	0.066

Table 4. Structural properties of the anonymized graphs generated for the synthetic graph 345 nodes.

	Original graph	Zhou and Pei	"AnonSN"
Context	(345,355)	(345,374)	(359,374)
Nber added edges	/	19	19
Number of false vertices added	/	/	14
APL	5.061038	4.750640	5.007765
Diameter	14	13	14
Radius	7	7	7
Medium degree	2.058	2.168	2.084
Density	0.006	0.006	0.006

We can notice from the tables that the value of the APL property calculated for the anonymized social graph using our "AnonSN" tool is always closer to its calculated value for the original social graph, so it is preserved. In the histogram below, we illustrate the values of this property (Fig. 9).

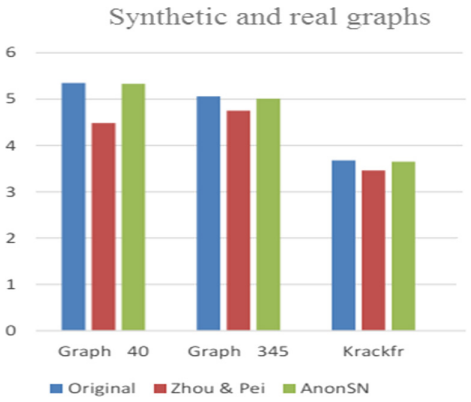

Fig. 9. APL for different data sets studied with k = 2

6.2 Analysis of the Variation of APL According to k

In this section, we study the performance of the proposed approach compared to the Zhou and Pei reference model as a function of different values of the anonymity parameter "k". Figure 10 shows the process of changing the value of APL of the graph "Interest_434" according to different values of k.

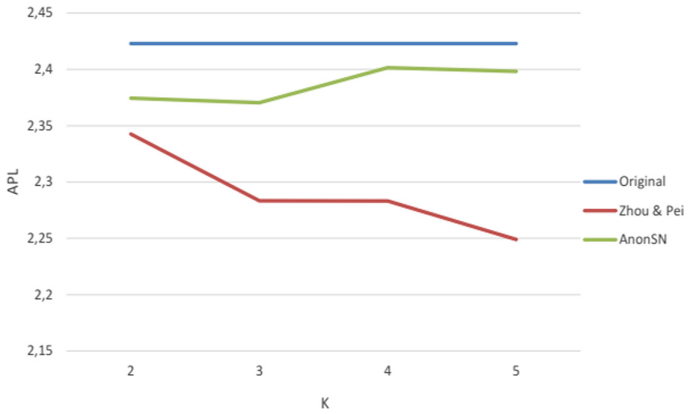

Fig. 10. APL values as a function of k

7 Conclusion and Outlook

In this paper, we have presented and analyzed the results of our proposed "AnonSN" anonymization tool, which considers the distance between nodes compared to the Zhou and Pei algorithm [1]. We have shown that our tool gives satisfactory results according to the tests carried out,it makes it possible to remove the changes in the distances between

the nodes and thus better preserve the APL structural property of the anonymized graphs, which will be more useful for data analysis. Our different experimental results demonstrate that the algorithm combining the addition of false nodes with the addition of links can obtain better results compared to that based only on the addition of links and it can generate a graph that effectively preserves the property APL.

1. The results indicate that our algorithm preserves the mean distances compared to the existing algorithm, because we add edges between the closest nodes, and if this addition does not preserve the mean distance, we add false nodes that maintain these distances close to those of the original graph.
2. Measurement of APL: We measured APL according to different values of the anonymity parameter k to confirm whether anonymization could avoid degrading the accuracy of analyzes by changing the distance between nodes. Specifically, we aim to maintain the value of the APL in an anonymized graph by comparing it to its original social network graph before anonymization as this preserves the usefulness of the data for future analysis, and we were able to achieve this goal.

At the end of this work, future extensions and perspectives are envisaged to improve it, namely:

- Deal with the case of $d > 1$, i.e. when the opponent has basic knowledge about the neighbors of the victim at d jumps.
- It would be interesting to reduce the number of added false nodes, in other words to study the number of added false vertices compared to:
  - to the total original number of vertices
  - the nature of the social graph to anonymize
- We also plan to study other network structural properties to further improve the preservation of utility in social network anonymization.
- Another interesting direction is to consider the implementation of this model in a graph with sensitive labels by introducing the concept of "l-diversity" to prevent homogeneity attacks and better protect sensitive labels while preserving structural properties.

References

1. Zhou, B., Pei, J.: Preserving privacy in social networks against neighborhood attacks. In: 2008 IEEE 24th International Conference on Data Engineering, pp. 506–515. IEEE (2008)
2. Tripathy, B.K., Panda, G.K.: A new approach to manage security against neighborhood attacks in social networks. In: 2010 International Conference on Advances in Social Networks Analysis and Mining, pp. 264–269. IEEE (2010)
3. Zhou, B., Pei, J.: The k-anonymity and l-diversity approaches for privacy preservation in social networks against neighborhood attacks. Knowl. Inf. Syst. **28**(1), 47–77 (2011)

4. Lan, L., Jin, H., Lu, Y.: Personalized anonymity in social networks data publication. In: 2011 IEEE International Conference on Computer Science and Automation Engineering, vol. 1, pp. 479–482. IEEE (2011)
5. Sweeney, L.: K-anonymity: a model for protecting privacy. Int. J. Uncertaintly Fuzziness Knowl.-Based Syst. **10**(05), 557–570 (2002)
6. Wu, H., Zhang, J., Wang, B., Yang, J., Sun, B.: (d,k)-anonymity for social networks publication against neighborhood attacks. J. Convergence Inf. Technol. JCIT **8**(2), 59–67 (2013)
7. Ghesmoune, M.: Anonymisation de réseaux sociaux. Ph.D. thesis, INRIA-IRISA Rennes Bretagne Atlantique, équipe S4 (2012)
8. Milgram, S.: The small world problem. Psychol. Today **2**(1), 60–67 (1967)
9. Faloutsos, M., Faloutsos, P., Faloutsos, C.: On power-law relationships of the internet topology. ACM SIGCOMM Comput. Commun. Rev. **29**(4), 251–262 (1999)
10. Wasserman, S., Faust, K.: Social Network Analysis: Methods and Applications. Cambridge University Press (1994)
11. Hay, M., Miklau, G., Jensen, D., Weis, P., Srivastava, S.: Anonymizing social networks. University of Massachusetts Amherst, Technical Report No. 07-19 (2007)
12. Bensimessaoud, S., Badache, N., Benmeziane, S., Djellalbia, A., et al.: An enhanced approach to preserving privacy in social network data publishing. In: 2016 11th International Conference for Internet Technology and Secured Transactions (ICITST), pp. 80–85. IEEE (2016)