# On the Accuracy Evaluation of Access Control Policies in a Social Network

Jedidiah Yanez-Sierra
*CINVESTAV - Tamaulipas*
Cd. Victoria, Mexico
jedidiah.yanez@cinvestav.mx

Arturo Diaz-Perez
*CINVESTAV - Guadalajara*
Guadalajara, Mexico
adiaz@cinvestav.mx

Victor Sosa-Sosa
*CINVESTAV - Tamaulipas*
Cd. Victoria, Mexico
vjsosa@cinvestav.mx

***Abstract*—Access control policies are mandatory for organizations whose operation involves sharing resources that must be kept private. In this paper, we address the problem of evaluating the accuracy of access control policies distributed in an interaction network modeled from a social network. Such network models granted access to documents owned by a large set of users. Since denied accesses are not included in the input network, we discuss a method based on Network Science to include complementary edges to have an approximate evaluation of ACPs' accuracy. The synthetic interactions allow the evaluation of ACPs by assessing the explicit and implicit intentions of the owners. We present an evaluation strategy to measure the accuracy of the generated ACPs. The results will be of interest to academics who want to synthesize information for similar phenomena.**

***Index Terms*—Access Control Policies, Complex Networks, Synthesize Information**

## I. Introduction

Two critical types of failures related to access control decisions are accesses that should be denied but are not or accesses that should be allowed but are not. Naturally, the consequences of these failures have different impacts on the security and privacy of the information [1, 2].

When users or organizations need to guarantee the security and privacy of their information, it is essential to identify and reduce possible failures related to low restrictive or highly over-adjusted access control policies (ACPs) [3, 4].

From a social network the set of interactions between a large set of users sharing information, complex network techniques can be applied to model a network of interactions, where vertices depict users (owners and consumers), and edges depict the explicit permissions between owners and consumers to interact with their content. In this scenario, the ACP assigned to that interaction can: (a) preserve or (b) break the accesses.

However, in an interaction-based graph, as well as in many other social network phenomena, only one type or part of the information is modeled or available. In the above scenario, only explicit access permissions are reflected. Therefore, it is necessary to define a mechanism that allows enhancing the information, to evaluate whether the explicit permissions continue to be preserved and if the implicit permissions are kept.

There is a need for synthetic data generation methods to perform proper inferences from them. Particularly, they play a special role when potential disclosure restricts the availability of the original data. Data collections have been used to produce synthetic versions of datasets when no real information is available.

In 1993 Rubin proposed a multiple imputation framework for synthetic data. Further contributions by Raghunathan et al. [5], provided a detailed methodology for making inferences from synthetic data. The authors simulated multiple copies of the population and release a random sample from each of these synthetic populations. Each synthetic dataset depicts the target population based on the collected data. Similarly, Penny et al. [6] evaluate the use of hierarchical Bayes imputation models for creating synthetic categorical data.

Benedetto et al. [7], in partnership with the U.S. Census Bureau, report the creation of a partially synthetic Census Bureau data product called the SIPP Synthetic Beta (SSB). The SSB has been extensively tested, looking for analytic validity over the years as new versions have been released. Similarly, Snoke et al. [8] evaluate and recommend methods to judge whether synthetic data have a distribution that is comparable to that of the original data. They also evaluate the extension of existing global and specific measures of utility and perform comparisons for data generated by different methods of synthesis.

As can be observed, works in the literature focuses on generating synthetic datasets that accurately model the original data, but there is a gap when trying to complete a dataset modeling the opposite behavior to the real one.

In this paper, we present a methodology based on Network Science to comprehensively address the problem of evaluating the accuracy of access control policies distributed in an interaction network modeled from a social network, which indicates explicit access to documents owned by a large set of users. Since denied accesses are not included in the input network, we discuss a method based on Network Science to include complementary edges in order to have an approximate evaluation of ACPs' accuracy. Our method takes advantage of the underlying information discovered when modeling the interactions between the documents and the users as a complex network.

We have tested our approach with a real dataset from the Instagram social network. Evaluating the accuracy of the ACPs associated with the users, the results showed that the proposed

978-1-7281-7624-6/20/$31.00 ©2020 IEEE
DOI 10.1109/CSCI51800.2020.00048

method allows us to adequately limit the number of synthetic edges to be added to the graph. Furthermore, it is shown that the synthetic edges adequately depict the opposite meaning to the real edges since they had an impact of up to 4% when considering the synthetic edges.

In Section II we provide the background of the proposed method. Section III describes the proposed methodology to generate synthetic interactions. Section IV describes the proposed evaluation methodology. Section V describes the experimental evaluation and the achieved results. Finally, some conclusions are discussed in Section VI.

## II. Background

Complex networks show properties that only emerge when modeling real massive phenomena, and in some human-made systems such as the internet network. Complex networks have been studied due to the particular characteristics they present in comparison to other types of networks. Characteristics as small-world effect, clustering, degree distribution, community structure, have been widely used to understand the modeled phenomenon [9].

An interaction network is a graph $G(V_G, E_G)$ that models a large set of interactions between documents and the users who use them. The set of vertices depict users (owners and consumers), and the edges depict the explicit permissions between owners and consumers to interact with their content. The interaction network can be processed through complex network techniques to generate, based on the underlying information its partition from two axes, horizontal and vertical. Horizontal partitioning groups users into communities and sub-communities, while a hierarchy of consumers and owners, depicted by $k$-shells, is created on the vertical axis.

Each vertex in the interaction graph has an associated ACP, which defines the consumers who can access the content of an owner. By using the interaction graph, the accuracy of the set of associated ACPs can be evaluated by analyzing whether a consumer fulfills the policy of a producer. However, only explicit accesses are covered in the graph, limiting the accuracy evaluation.

### A. *Definition of the Accuracy Metric*

In binary classification, data is divided into two different classes, positives (P) and negatives (N). The binary model then classifies the input instances as positive or negative. Given the inputs, the model outputs (classification), and the actual (reference) outputs, 4 groups of results are identified; 2 types of correct classification (true), true positive (TP) and true negative (TN); and two types of incorrect classification, false positive (FP) and false negative (FN) [10].

A 2x2 table formulated with these four outcomes is called a confusion matrix, which is used to describe the performance of a classification model in a dataset for which both true values and the resulting values from the model are known [11, 12]. Fig. 1 shows a typical confusion matrix. Cells in green represent correct model predictions, while red ones represent incorrect predictions.

| | | **Predicted/Classified Class** | |
|---|---|---|---|
| | | *Negative* | *Positive* |
| **Actual** | *Negative* | True Negative (TN) | False Positive (FP) |
| | *Positive* | False Negative (FN) | True Positive (TP) |

Fig. 1: Illustrative example of a binary confusion matrix.

Once defined the positive and negative possible outcomes, the accuracy can be defined as the ratio of the correctly classified inputs to the whole dataset. Formally, the accuracy is defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

## III. Method to Complement the User Interactions

Given the lack of real datasets having all the necessary cases to evaluate social phenomena, here we describe the proposed methodology based on network science to synthesize complementary interactions that model implicit denied access in a network of interactions.

The proposed methodology consists of two steps, synthesis of a large set of possible complementary interactions followed by sampling a representative set of them. The set of ACPs associated with the users must preserve the real accesses; however, given the nature of the information, there is a possibility that ACPs do not fulfill the original authorizations of the owners. The ACPs could: preserve or break the real interactions.

Furthermore, in an interaction-based graph, only explicit access permissions are reflected. Therefore, it is necessary to define a mechanism that allows enhancing the information, not only to evaluate whether the explicit permissions continue to be preserved (TP, FN) but also if the implicit permissions (TN, FP) are kept.

As a result of evaluating the ACPs, there are only two possible outputs, the policy allows or denies access. TP and FN are two cases implicit in the graph. To include the two remaining cases, TN and FP, it would be necessary to evaluate the set $ES_G$ of all the possible combinations of edges not explicitly present in the set $E_G$, where $ES_G \rightarrow \{\{v, w\} \notin E_G, |\, v, w \in V_G\}$.

Since the interaction graph behaves like a complex network, it is sparse. Only those complementary edges (also called synthetic) in the scope of users and having meaningful characteristics need to be included. In addition, the set of possible missing interactions is large enough and only a small subset of them, lower than the real number of interactions, needs to be incorporated in an accuracy analysis, otherwise, the real interactions would be hidden.

### A. *Construction Based on Graph Partitioning*

Partitions in the graph are used to generate synthetic edges that imitate the behavior of the original ones. That is, 1) consider only vertices within the same community or sub-community, 2) preserve the hierarchy. Only edges from lower shells to upper ones would be considered, as well as only vertices in the same partition.

We assume that, since users are ranked from consumers to producers, from lower $k$-shell to higher $k$-shell, it would be unlikely the existence of synthetic edges whose source is a vertex in an upper $k$-shell towards nodes in lower $k$-shells.

### B. Filtering Based on Similarity

The selection of possible synthetic edges to be added to the set $ES_G$ can be further refined by applying strategies used in the link identification problem. The Resource Allocation Index (RAI) is defined as the number of resources a node $w$ receives from a node $v$ through indirect neighbors [13].

Let $\delta_i$ be the set of nodes adjacent to vertex $i$, the resource allocation index is defined as,

$$SI_{RA}(v,w) = \sum_{z \in \delta_v \cap \delta_w} \frac{1}{k(z)}, \tag{2}$$

where $k(i)$ denotes the degree of $i$. To follow the topology of the interactions (from lower $k$-shells to upper ones), only the outgoing edges are taken from the neighborhood of vertex $v$, and for the neighborhood of node $w$, only the incoming edges are considered.

RAI with the above modification is used to compute the similarity between pairs of nodes. For each pair of vertices, the similarity is associated with the possible synthetic link formed between them. High similarity values depict edges with similar meaning as the real ones. But our goal is to add edges with the opposite meaning, so instead of considering the higher similarity values, the lower ones are used. However, the lower similarity range can have a bias, also called flattening, associated with the neighborhood wideness used in the similarity metric. Nevertheless, since users are grouped into communities and sub-communities, neighborhoods cannot extend beyond these.

### C. Sampling Based on a Probability Distribution

Using only vertex similarity to filter possible edges could generate to many cases. To overcome this, an edge-sampling procedure based on a probability distribution function $F_X$ is being used to select a random subset, $E'_G$, of $ES_G$.

The Algorithm 1 describes the complete process to generate the edges for the set $E'_G$ that will complement the edges of the graph. The algorithm receives as input the interaction graph $G(V_G, E_G)$, a function associated with a probability distribution $F_X$ and a probability of selection $P_{E'}$. The function $F_X$ and the probability $P_{E'}$ are used in the sampling process to control the number of edges that will be added to $E'_G$. As output, the Algorithm 1 returns the augmented graph $G^+(V_G, E_G \cup E'_G)$, which includes the synthetic edges; those will be considered the TN and FP cases.

**Algorithm 1:** Process to generate the synthetic edges.

**Data:** $G(V_G, E_G)$: Graph of interactions
$F_X$: Probability Distribution Function
$P_{E'}$: Selection Probability
**Result:** $G^+$: Graph complemented with the synthetic edges

$E'_G \leftarrow \{\}$
$C \leftarrow getCommunities(G)$
**foreach** $c \in C$ **do**
  $G_C \leftarrow subgraph(G, c)$
  $SC \leftarrow getCommunities(G_C)$
  **foreach** $sc \in SC$ **do**
    $G_{SC} \leftarrow subgraph(G_C, sc)$
    $KS \leftarrow getOrderedKS(G_{SC})$
    **foreach** $k \in KS$ **do**
      $V_{ks} \leftarrow v \in V_{G_{SC}} \mid ks(v) == k$
      **foreach** $v \in V_{ks}$ **do**
        $W_{ks} \leftarrow w \in V_{G_{SC}} \mid ks(w) > k \cup \{v, w\} \notin E_G$
        **foreach** $w \in W_{ks}$ **do**
          $s \leftarrow Similarity(v, w)$
          $p \leftarrow P(v, w, s, F_X)$
          **if** $p \leq P_{E'}$ **then**
            $E'_G \leftarrow E'_G \cup \{v, e\}$
          **end**
        **end**
      **end**
    **end**
  **end**
**end**
**return** $G^+(V_G, E_G \cup E')$

## IV. Measuring Policy Accuracy

### A. Simple Node-Based Strategy

The node-based evaluation process is simple and straightforward. For each vertex $v$, its output neighborhood $\delta_v^{out}$ is evaluated. For each neighbor $w \in \delta_v^{out}$, the result of the ACP is evaluated, obtaining the proportion of real edges that are TP and FN, and the proportion of synthetic edges which are TN and FP. Based on the edge type, real or synthetic, and the result of $v$'s policy validation against the $w$'s one, the edge is counted in one of the four possible cases: 1) real edges and positive access into TP, 2) real edges and denied access into FN, 3) synthetic edges and positive access into FP, 4) synthetic edge and denied access into TN. Algorithm 2 shows the complete process to compute the node-based evaluation. The accuracy value is calculated based on the Definition 1.

### B. Edge-Weighted Node Strategy

Each real edge in the interaction graph may have an associated weight (interaction weight). The greater the relationship between a pair of users, the greater their weight. The interaction weight can be used to weigh the positive or negative cost due to an ACP. Then, the impact of each relationship will be linked to its weight, thus having weighted TP (wTP) and weighted FN (wFN).

Edge weights can have different meanings, such as quality, importance, strength, security level, etc., then using edge weights can help to enhance the meaning of the accuracy value. However, it is necessary to introduce weights to the synthetic edges for the FP or TN type of interactions.

*1) Assigning weights to synthetic edges:* To compute the weights for the synthetic edges, two strategies are proposed. Let $z \leftarrow SI_{v,w}$ be the set of edges' weights linking the nodes that are in the intersection of $v$ and $w$ neighborhoods. If $z$ is not empty, the weight is obtained by a random number with normal distribution with parameters $\mu_z$ and $\sigma_z$, where $\mu_z$ and $\sigma_z$ are the mean and the standard deviation of the weights associated with the edges in $z$. If $z$ is empty, the

**Algorithm 2:** Node-based Evaluation Process.

**Data:** $G(V_G, E_G \cup E'_G)$: Graph of interactions
**Result:** $accuracy$: Accuracy value for the ACPS assigned to the graph

1 $TP, TN, FP, FN \leftarrow 0$
2 **foreach** $v \in V_G$ **do**
3     **foreach** $w \in \delta_v^{out}$ **do**
4         **if** $\{v, w\} \in E_G$ **then**
5             **if** $ACP_v$ **satisfies** $ACP_w$ **then**
6                 $TP \leftarrow TP + (1/|\delta_v^{out}|)$
7             **else**
8                 $FN \leftarrow FN + (1/|\delta_v^{out}|)$
9             **end**
10         **else if** $\{v, w\} \in E'_G$ **then**
11             **if** $ACP_v$ **satisfies** $ACP_w$ **then**
12                 $FP \leftarrow FP + (1/|\delta_v^{out}|)$
13             **else**
14                 $TN \leftarrow TN + (1/|\delta_v^{out}|)$
15             **end**
16         **end**
17     **end**
18 **end**
19 $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$
20 **return** $accuracy$

weight is obtained using the mean and standard deviation from the weights associated to the set in the intersection of outgoing edges of $v$ and incoming edges of $w$.

*2) Using the edge weights:* The node-based and edge-based approaches can be reformulated to consider weights. In the node-based weighted approach, lines 6, 8, 12, and 14 of the Algorithm 2 are replaced by using eq. 3, where $X$ stands for one of the four cases wTP, wFN, wTN, or wFP.

Equation 4 shows the function to calculate the total weight of the outgoing neighborhood of node $v$, value that is needed to calculate the proportion of weight for each interaction. Once the total weight is calculated, the values of wTP, wFN, wFP, and wTN would be calculated as defined in eq. 3,

$$X \leftarrow X + weight(v, w)/W_{tot}^v \tag{3}$$

$$W_{tot}^v \leftarrow \sum_{w \in \delta_v^{out}} weight(v, w) \tag{4}$$

where $v$ is a node, $\delta_v^{out}$ is the set of outgoing neighbors of $v$, $X$ may be one of the confusion matrix values, and $weight(v, w)$ is the weight of the edge from $v$ to $w$.

### C. Simple Edge-Based Strategy

Unlike the node-based method, in the edge-based approach, all the edges are considered individually, evaluating the source and target vertices.

All edges of the augmented graph, $G^+$, are evaluated to generate the values of the confusion matrix, according to the type of edge (real or synthetic) and the ACP validation of the adjacent vertices.

ACPs assigned to vertices joint by real edges should preserve access (TP). If the ACP denies access, the edge is counted as FN. On the contrary, all ACPs assigned to vertices joint by synthetic edges should deny access and must be counted as TN. Otherwise, they would be FP.

### D. Edge-Weighted Based Strategy

Edges weights can be used to refine the quality of the accuracy metric. Simple counters can be replaced with operations such as $X \leftarrow X + weight(v, w)$, where $X$ can be any of the four values of the confusion matrix, and $weight(v, w)$ is the weight associated with the edge joining $v$ and $w$. In this way, each interaction will be weighted by its associated weight.

## V. EVALUATION

We tested our proposal using a dataset of real users and their interactions. The dataset was modeled as a complex network, which was analyzed using complex network techniques to deliver a graph with the partitioning this methodology requires. Also, each user has been assigned an ACP.

We build a prototype that performs all tasks presented in Section III. The prototype was coded in *python* and *igraph* library.

### A. Dataset Description

The dataset consists of 17K Instagram users having more than 173K interactions. The dataset was crawled through the Instagram API by Ferrara et al. [14]. It contains anonymized public media and user information from the social media Instagram.com.

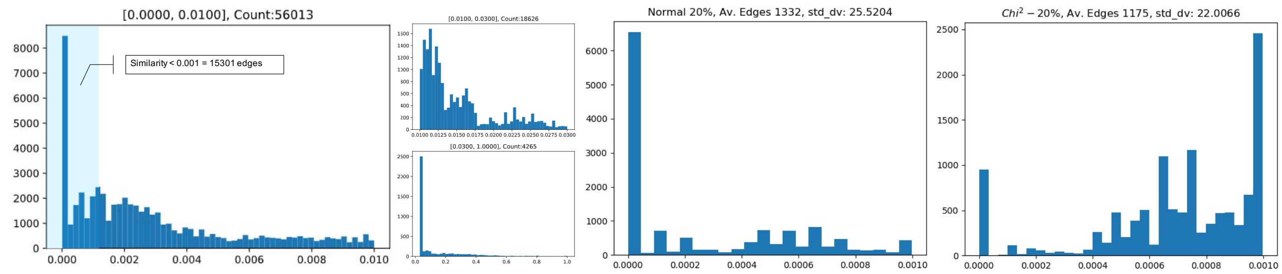
The full graph has 173K edges, Avg., Path Length: 3.44, Cluster Coefficient: 0.11, Diameter: 12, and Assortativity Index: 0.101. As can be seen, the graph fulfills the properties of a complex network despite the relatively low cluster coefficient. The vertices of the graph were grouped into 96 communities which are partitioned into 312 sub-communities and 237 $k$-shells. 54% of the communities are further partitioned into 2 to 28 different sub-communities. The average number of sub-communities per community is 10.81.

For simplicity, here we show the results obtained for community 23, the largest and most representative community of the network. Its main characteristics are: 3150 vertex, 23K edges, Average Path Length 2.83, Cluster Coefficient 0.183, and Diameter 6.

### B. Results of the filtering based on similarity

After applying the method to generate synthetic edges based on graph partitioning, only 78,904 synthetic edges were generated from the 4.9 million possible for the community 23. Fig. 2a shows the histogram with the similarity distribution for the set of synthetic edges. The normalized similarity values are shown on the horizontal axis. Each image shows a different range, the left image shows the range $[0, 0.01)$ having 56,013 edges, the upper-right figure shows the range $[0.01, 0.03)$ with 18,626 edges, and the lower-right shows the range $[0.03, 1]$ with the remaining 4,265 edges. The vertical axis shows the number of edges for each bucket.

In the left graph of Fig. 2a, it can be observed the flattening effect for the lowest similarity values having more than 8K vertices. Further, in the range $[0, 0.001]$ there are 15,301 vertices, which is just over twice the real interactions in that community. Therefore, according to the second consideration in generating




(a) Histogram of similarity values of the complete set of generated synthetic edges.

(b) Similarity values of edges sampled with normal distribution.

(c) Similarity values of edges sampled with $\chi^2$ distribution.

Fig. 2: Distribution histograms of similarity values assigned to the generated synthetic edges. Sampled histograms are the average of 32 executions of the sampling schemes using a proportion of 20% regarding the real edges.

synthetic edges, only the edges in the range $[0, 0.001]$ are considered. This range varies according to community sizes.

### C. Results of the sampling based on a probability distribution

Two different probability distribution functions were tested: *normal* and *chi-square* ($\chi^2$). The uniform distribution was selected as the base case since its density function is constant for all elements in the range, i.e., all the synthetic edges would have the same probability of being selected. On the other hand, chi-square ($\chi^2$) distribution was selected due to the shape of their density function, assigning a higher probability to edges with a lower similarity value, see Fig. 2. The two different probability functions illustrate different ways of edge-sampling, and here are used to show the accuracy loss trend while adding synthetic edges.

The input parameters to shape the probability distributions are selected in such a way that it extended in the selected similarity range ($[0.0, 0.001]$ for community 23). For the uniform distribution, the range of similarity is equivalent to the range defined for the distribution, i.e., $[0.0, 0.001)$. For the chi-square distributions ($\chi^2$), were used the values, $k = 3$ and a scale $1\times10^{-3}$. For each probability distribution, Fig. 2 shows the histogram with the average number of sampled edges.

As can be observed in Fig. 2, by using a uniform distribution, all synthetic edges have the same probability of being selected. By using the chi-squared distribution, the range selected is skewed towards the upper range of similarity.

### D. Results Using Both Real and Synthetic Edges

Since the information in $E'_G$ is generated synthetically, for the purpose of this work, its acceptable range is limited to $0 \leq |E'_G| \leq |E_G|$, thus maintaining a balance between the amount of real data and the number of synthetics.

By introducing more synthetic edges, accuracy is decreased from the value obtained only with real edges. The different sampling by a distribution function also affects the accuracy.

Fig. 3 shows the results when evaluating accuracy for different proportions (in the range $0 \leq |E'_G| \leq |E_G|$) of synthetic edges. The figure resumes the accuracy results for the four evaluation schemes described in Section IV.

The proportion of edges added to the set $E'_G$ is shown on the horizontal axis of Fig. 3. The left end means no synthetic edges added and the right end (1.0) means as many synthetic edges added as the number of edges in the original graph. Each point in the graphs shows the average accuracy values obtained for 32 runs for each proportion test of synthetic edges added.

All the results in Fig. 3 start in the base case, i.e., when no synthetic edges have been added, and from there, a reduction is observed as more synthetic edges are added. As can be seen in the figure, the base values are 92.11 and 93.28, for the simple and weighted vertex-based assessment, and 87.21 and 83.88 for the simple and weighted edge-based evaluation. In the four cases, from the base value, low loss in accuracy is observed, which is more pronounced at the beginning (synthetic edges proportions $< 50\%$) and slightly smoother at the end (proportions $> 50\%$).

The difference between including or not the weights in the vertex-based evaluation has a positive impact of 1.17% in the base value and only 0.18% when considering a proportion of 100% and chi-square distribution. For the edge-based evaluation, the impact is also positive on both ends, having 3.33% in the base value and 1.0% on the other side and the chi-square distribution.

It can be observed that the accuracy loss trend is more or less the same in any case. Adding a number of synthetic edges greater than 100% of original edges would introduce too much noise to the information since synthetic data cannot fully replicate real data, especially if it tries to replicate non-existent behaviors; therefore, the more synthetic data is used, the less relevant the data could be.

To preserve an adequate balance between the source set and the synthetic data, a reasonable proportion of synthetic edges would be less than 30%. This range will allow the evaluation of the implicit accesses of the owners (unauthorized access), without detracting from the quality of the information.

Figure 4 shows the behavior of the four variables of the confusion matrix. As can be seen, in the base case, there are no TN or FP values. It can be seen in the figure that the proportion of FN on average is kept for all cases, while the TP value decreases as the TN and FP values increase, indicating that the ACPs have some FP-type faults, which in turn affects the accuracy.

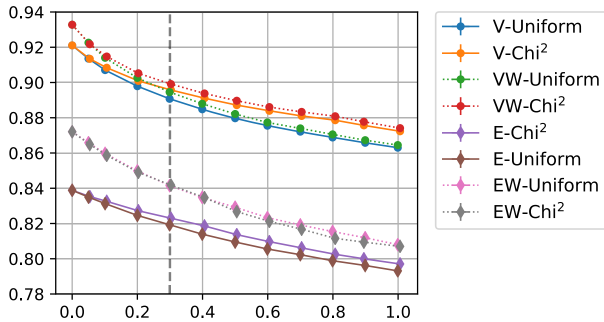



Fig. 3: Results of the Accuracy evaluation using the four evaluation approaches and the three random filtering schemes for the complete interaction graph $G^+$. Values are the average of 32 independent runs. Vertical line indicates a suitable value for the proportion of synthetic edges.

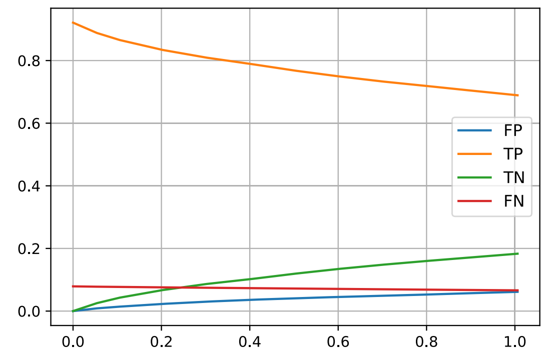



Fig. 4: Average behavior of TP, FP, TN, FN values when evaluating the network $G^+$ using chi-square sampling function.

Based on the results, it can be concluded that the synthetic edges adequately generate both FN and FP cases. The synthetic edges allowed to prove that in addition to the FN cases identified in the ACPs, there is also a portion of ACPs that allow unauthorized access (FP), which contributed to the accuracy decrease.

## VI. Conclusions and Future Work

In this paper, we presented an approach based on complex networks to take advantage of the underlying information discovered to generate complementary interactions in an interaction network modeled from a social network. In our study case, the set of interactions models the accesses allowed between a large set of users that share an equally large set of documents. However, the method can easily be applied to complement the information of other phenomena modeled equivalently. The synthetic edges allowed the evaluation of both explicit and implicit actions of the owners.

Achieved results show that our proposal properly evaluates the policies that had been assigned to the users, showing that the synthetic interactions in effect depicted an opposing behavior regarding the real interactions. If the policies were error-free, the values for the synthetic edges would have been distributed only in TN values, which would not affect the accuracy value. However, the observed reduction shows that in addition to FN, the policies allow some FP, which can have a great impact on security and access control scenarios.

We are testing our scheme with other similar datasets, and also, we are exploring a general model to complement the information in the graphs that can be applied to different phenomena.

## References


[1] M. Beckerle and L. A. Martucci, "Formal definitions for usable access control rule sets from goals to metrics," *SOUPS 2013 - Proc. 9th Symp. Usable Priv. Secur.*, 2013.

[2] M. Mondal, M. T. Khan, C. Kanich, G. S. Yılmaz, M. Tang, B. Ur, N. Hirsch, C. Tran, and E. Zheleva, "Moving beyond set-it-and-forget-it privacy settings on social media," *Proceedings of the ACM Conference on Computer and Communications Security*, pp. 991–1008, 2019.

[3] S. Egelman, A. Oates, and S. Krishnamurthi, "Oops, I did it again: Mitigating repeated access control errors on facebook," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 2295–2304, 2011.

[4] A. Nadkarni, W. Enck, S. Jha, and J. Staddon, "Policy by Example: An Approach for Security Policy Specification," jul 2017. [Online]. Available: http://arxiv.org/abs/1707.03967

[5] T. Raghunathan, J. Reiter, and D. Rubin, "Multiple Imputation for Statistical Disclosure Limitation," *Journal of official statistics*, vol. 19, no. 1, p. 1, 2003.

[6] R. Penny, P. Graham, and J. Young, *Methods for Creating Synthetic Data*, 2008, vol. 3.

[7] G. Benedetto, M. H. Stinson, and J. M. Abowd, "The Creation and Use of the SIPP Synthetic Beta," no. April, pp. 1–19, 2013.

[8] J. Snoke, G. M. Raab, B. Nowok, C. Dibben, and A. Slavkovic, "General and specific utility measures for synthetic data," *Journal of the Royal Statistical Society*, vol. 181, no. 3, pp. 663–688, 2018.

[9] A. Garcia-Robledo, A. Diaz-Perez, and G. Morales-Luna, "Characterization and Coarsening of Autonomous System Networks:," in *Adv. Methods Complex Netw. Anal.* IGI Global, 2016, vol. i, pp. 148–179.

[10] L. Derczynski, "Complementarity, F-score, and NLP evaluation," *Proceedings of the 10th International LRE Conference, LREC 2016*, pp. 261–266, 2016.

[11] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLoS ONE*, vol. 10, no. 3, pp. 1–21, 2015.

[12] A. Tharwat, "Classification assessment methods," *Applied Computing and Informatics*, 2018.

[13] M. R. Srilatha P, "Similarity Index based Link Prediction Algorithms in Social Networks: A Survey," *J. Telecommun. Inf. Technol.*, vol. 2, no. 2, pp. 87–94, 2016.

[14] E. Ferrara, R. Interdonato, and A. Tagarelli, "Online popularity and topical interests through the lens of instagram," in *Proc. 25th ACM Conf. Hypertext Soc. media - HT '14*, no. i. New York, New York, USA: ACM Press, 2014, pp. 24–34.