Global-Structure Aware Sub-Graph Classification for Identical Author Detection

Mahi Rahimi
School of EE
KAIST
Daejeon, South Korea
mahi@kaist.ac.kr

Rodrigo Hormazabal
Graduate School of AI
KAIST
Daejeon, South Korea
rshormazabal@kaist.ac.kr

Jeongeon Park
School of Computing
KAIST
Daejeon, South Korea
jeongeon.park@kaist.ac.kr

ABSTRACT

In the online world where identities can be easily changed, often multiple identities are created for a single person. In order to prevent situations where the duplicate identities are misused, they need to be correctly identified. In this project, we use a dataset composed of author IDs and publication IDs, where a single author can have multiple IDs. To achieve this, we propose **GASC**, Global-Structure Aware Sub-Graph Classification approach where the information from the entire graph is used to detect duplicate authors. Through experiment, we show that our approach achieves a better accuracy than the baseline, and that the approach and be generalized to any randomly selected nodes in the graph.

1. INTRODUCTION

In the current age of internet and social media, a big part of people's online interactions are done using a pseudonym or just anonymously. As the virtual world becomes more inclusive, it is important to understand the relationship between user's virtual and real identity.

An important question regarding these interactions, looking to avoid a series of problems related to the misuse of this anonymity, is whether a set of multiple virtual identities belong to a unique person/agent or not. Understanding this problem not only helps to keep a safer and more consistent environment with reality, but can also be beneficial when we need to assign credit or blame a user in certain situations.

Specifically, in this project we work with a dataset comprised of authors and publications, in which one author might have more than one identity. We are given two different types of dataset including paper-author relationships and author-ID pairs. In order to correctly acknowledge an author's influence and correctly attribute the importance of its own work, finding out an author's different IDs and their publications is a fundamental step.

In most identity recognition settings, current approaches use users' information or the text information from the paper to find a mapping between their information and their identity [3, 4, 5]. However, the above approaches cannot be applied as no information other than the pairs and the IDs are provided. In this project, we tackle the problem from a different perspective: we find authors with multiple ID's only based only on the given paper-author relationships.

This approach is more general, meaning that we can transfer the problem to another application while being less computationally expensive, since we do not use semantic information contained in the sample itself, like in the field of natural language processing or computer vision.

However, after carefully exploring and analyzing the data, we claim that the problem is difficult conceptually because there might be a weak correlation between the structure of an author's neighborhood and the fact that both ID's correspond to the same individual. Although achieving high accuracy in validation set could be attributed to the selected distribution of validation set which is different from the real setup, where all pairs of nodes might be considered and the ratio between positive and negative labels is really low.

First, we implement and compare decision-based methods by considering the centrality and common co-authors as a measure of similarity between the authors. We compared these methods with a random classifier and show its performance as a baseline. Furthermore, we argue about the downside of this methods and poor performance in out of distribution data, which leads to proposing a learning scheme for covering this weakness.

With the exploration, we propose a sub-graph classification method with graph neural networks (GNN). The approach benefits from finding structural similarities between same authors in the graph. A bipartite graph between authors and papers used as the knowledge representation, and the details are presented in Section 2.5.

The contributions of this project are the following:

- We first propose some centrality measures which can classify the validation set with high accuracy.
- We argue that this accuracy can not be generalized to any test scenario, and propose GASC, Global-Structure Aware Sub-Graph Classification approach where only the author-paper relationships are taken into account in detecting duplicates.
- We show generalization of our method empirically not only on the performance of methods on validation set but also on any randomly selected nodes from the graph.

.

2. PROPOSED METHOD

2.1 Data Analysis

We analyzed the data in terms of number of author corresponding to each paper, similarity between co-authors, and centrality measures to the downstream task.

In the end, we visualized the results in the form of histograms.

1) Centrality and Common Co-authors We are trying to find how different are the distributions between duplicated and non duplicated authors regarding co-author similarity or centrality measures within the graph.

We use Fig.2(a,b) to approve the intuition behind our baseline and also the order of correlation which we use in implementation.

2) Histograms We show the distribution of authors to corresponding papers and other authors, this finding leads us to better understanding of our problem and steps we need to take to come up with appropriate representation of graph.

2.2 Baselines

As a control baseline for this project, several methods are suggested.

1) Random Decision (weak baseline) Randomly decide whether two author IDs belongs to same author or not. Also considering if one paper contain the both author ID, those IDs does not belong to same the author.

2) Centrality (medium baseline) Considering the centrality as a factor of similarity as long as there is no paper which both author ID appears.

3) Common Co-author (strong baseline) Considering the number of common co-author occurrences as a factor of similarity as long as there is no paper in which both author IDs appears. These intuitions comes from domain knowledge of how is the relationship between author and paper and might not work in other tasks.

We can also refer to Fig.2(a,b) to approve the intuition behind this baseline and also the order of correlation which we might use in implementation of baseline.

2.3 Graph Representation

The data is modeled as a bipartite graph with authorID and paper as two side of the graph. To reduce amount of information flow in decision making, for every pair of node under study we use the sub graph which contain all the path between two nodes with the length shorter than five. We use node degree and local profile as feature representation for each node, which is shown to be strong representation in graph classification tasks [1, 2].

2.4 Problem Formulation

Three different ways are suggested for the formulation of the graph classification.

1) Mean/Max Pooling After few round of message passing we apply a pooling for all the nodes in graph and train a classifier on top of pooled features.

2) Reduce all to a node Add one dummy node to the graph which is connected to all other nodes, not only improve the information flow between nodes [], but also can be used as the input of classifier.

3) Reduce two candidate to a node Add one dummy node to the graph which is only connected to two candidate author under examination, and apply classifier on dummy node, this representation lower the chance of false positive in case a same author appears in sub-graph beside candidates.

All representation reach to same performance, but *reduce all nodes to one node* have more stable training and faster convergence compared to other two.

2.5 GASC: Global-Structure Aware Sub-Graph Classification

GASC training scheme has three important parts which will be described in detail here.

1) Graph Model: Most of the information necessary to discriminate between a duplicated pair of authors is contained in the sub-graph made by their neighborhoods. Even though, totally disjoint neighborhoods could also represent a positive pair of authors (in the unlikely case that an author's publications is split in two sets of completely separate group of collaborators), its futile to to try to predict these cases since most of the information necessary to distinguish these situations wouldn't be contained in the author-paper interactions and extra features would be needed.

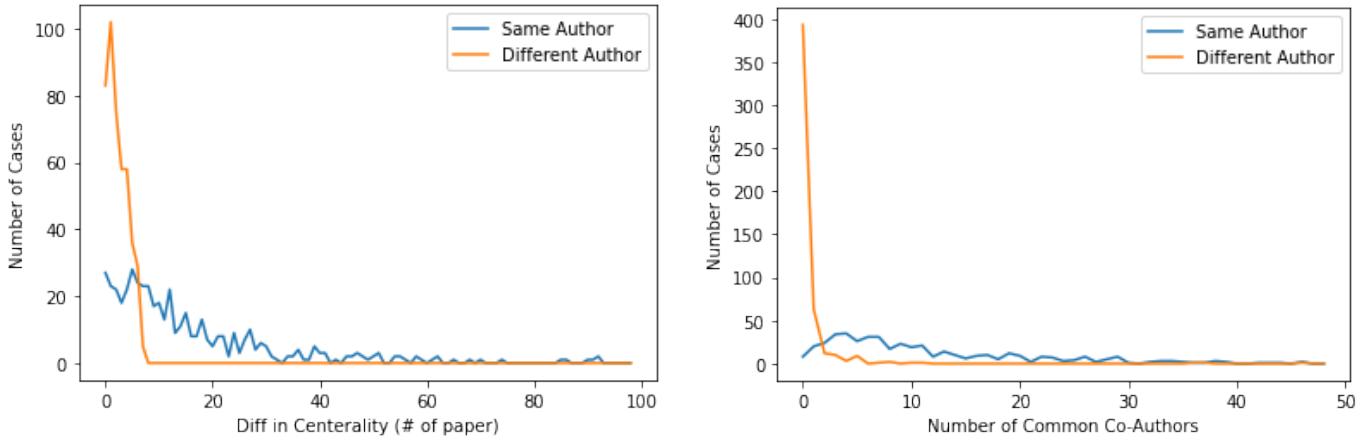
We tackle this problem by first encoding all nodes information about their neighborhood and features within the whole author-paper bipartite graph and then we create graph subset made of the nodes and edges in the k-hop path between a queried pair or nodes (to predict duplicated authors). After, creating the smaller graph containing all the paths information we run a simple graph neural network that tries to learn the small patterns and differences between subgraphs of author pairs. We do thus under the assumption that we can generalize a set of rules between subset graphs instances that will get to create meaningful representations for this pair of nodes within the sub-graph.

In the sub-graph we add one additional dummy node to increase information flow and also use the feature represent on extra node for base of classifier. The node features start with all zero.

2) Data Point Selections: In order to sample negative examples to learn an effective representation, its imperative to design an strategy that focus on finding author pairs that have the minimum right conditions of a duplicated pair. This will allow us to find complex patterns that can discriminate between duplicated pairs of nodes and cases were authors juat have a highly overlapping network of coauthors. First, duplicated authors must have at least one direct path between each other, which greatly reduces the space of possible positive pairs. Also, as mentioned earlier, authors with a co-authored paper can not be a duplicated pair and, cases with an small number of 4+hop paths to each other are far less likely to be duplicated. A natural strategy that arises from these restrictions is to take nodes in the training data and pair them with other authors in the training data that meet these criteria. Assuming authors can only be duplicated one time, we can be sure that samples pairs from this data cannot be a positive pair, since the only positive author pair its the corresponding labeled paired nodes.

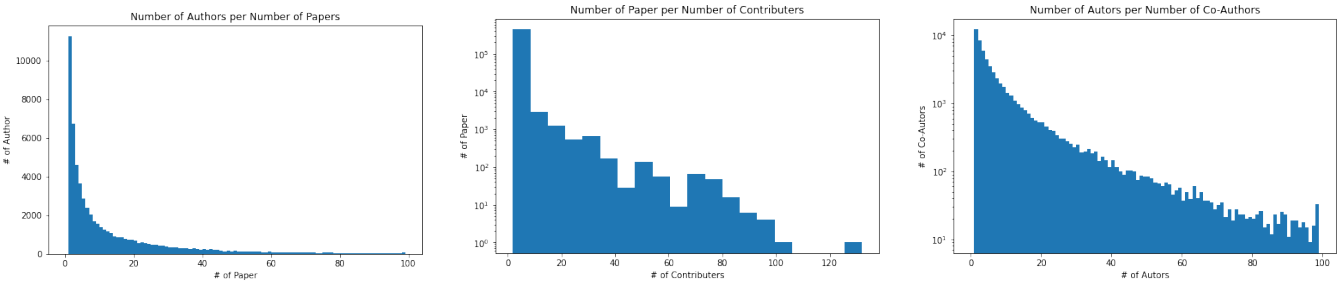
3) Neural Network Architecture: We focus on using two graph convolution layer and one fully connected layer as classifier on reduction node, the whole architecture is shown in **Figure** 3.

Varying graph convolution layer tried for this problem (GAT, GIN, GCN, SAGE) but GCN shows stable convergence while SAGE and GIN was faster in reaching converges



(a) (b)

Figure 1: Correlation between differences in centrality (a) and common co-author (b) to same author



(a) (b) (c)

Figure 2: Histograms of paper to author (a), author to paper (b), author to co-author (c)

point, and there was no significant different in final performance.

We trained the network with batch size of 1024 and 128 width of hidden layers.

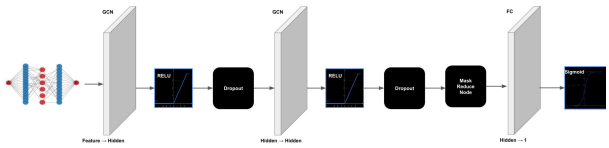


Figure 3: GASC's GNN Model Architecture

2.6 Evaluation Methods

For evaluation of our methods, we used several metrics. One metric cannot give a full understanding on what the weaknesses of our model are, so we used a combination of the following.

1. **Accuracy**: is the fraction of correctly predicted data point out of all data points. The problem is that we assign equal cost to false positives and false negatives when calculating it. So, even though accuracy is very intuitive, it does not work well for imbalanced classes.

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

2. **F1 Score**: is the harmonic mean of precision and recall. F1 score conveys the balance between the precision and the recall. In order to get high F1 score, the model should predict both predict a large fraction of the all tags, truly corresponding to this question, and a large fraction of the predicted tags should be truly relevant for the question.

$$\begin{aligned} F1\_Score &= 2 * \frac{Precision * Recall}{Precision + Recall} \\ &= \frac{2TP}{2TP + FP + FN} \end{aligned} \tag{2}$$

3. **Precision**: is the fraction of relevant instances among the retrieved instances. Intuitively, it means how many of our predicted tags are correct.

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

4. **Recall**: is the fraction of relevant instances that were retrieved. Intuitively, it means how many of the correct tags we actually predicted.

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

3. EXPERIMENTS

Beside validation set we also propose two other experiment setup which shows generalization of solution beyond valid set. Comparison of baseline and GNN model will be shown in below plots. The proposed methods based on GNN outperform other baselines in all three scenarios.

For better comparison we start by **Figure** 4 showing the evaluation with random decision-making, as the threshold increases we can see recall and f1-score, since half of validation data are positive cases we don't see any improvement in precision or accuracy.

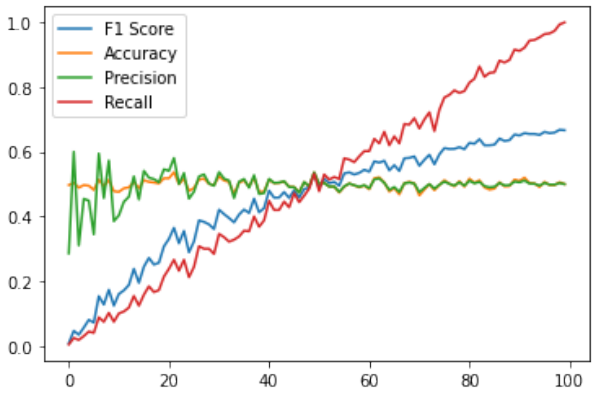


Figure 4: Random Baseline: (Accuracy: 0.537, F1-Score: 0.668, TH: 21)

Having this in mind we can now compare co-author baselines and GASC in **Figure** 5. The GASC is more stable with regard to threshold and outperforming our baseline by 10% raise in accuracy and F1-score.

For a better comparison between baselines and GASC, **Figure** 6 shows one parameter over different models.

The negative data point in validation are sampled from distribution with low density in sub-graphs unlike train data points. For second comparison we sample valid dataset completely random. There are 3000 positive cases in two billion total possible pairs, we label every random pair as negative since it is really unlikely to sample a positive case, therefore we add train set data again to random selected samples for evaluation of models. **Figure** 7

Another challenging case is to sample pairs with high density sub-graphs, since this pairs are more similar to train set and more likely to be false positive, the GASC can label them negatively as good as other samples when they sampled randomly while other baseline methods have significant drop in performance. **Figure** 8

4. CONCLUSIONS

Understanding and preprocessing the data is important task in data mining, although it will give us good information and intuitive idea, that might not be enough for building a general and adaptive solution.

To create an adaptive and general solution we need to use methods which can learn and improve upon new labels.

In this work, we introduce a new problem formulation based on k-hop sub-graph classification (GASC) which train on a very few positive samples to detect identical authors. We show the importance of carefully picking negative samples to reach a high accuracy and generalization.

At the end, we present result of GASC in compared to other baselines not only on validation set but also on two other benchmarks.

5. REFERENCES

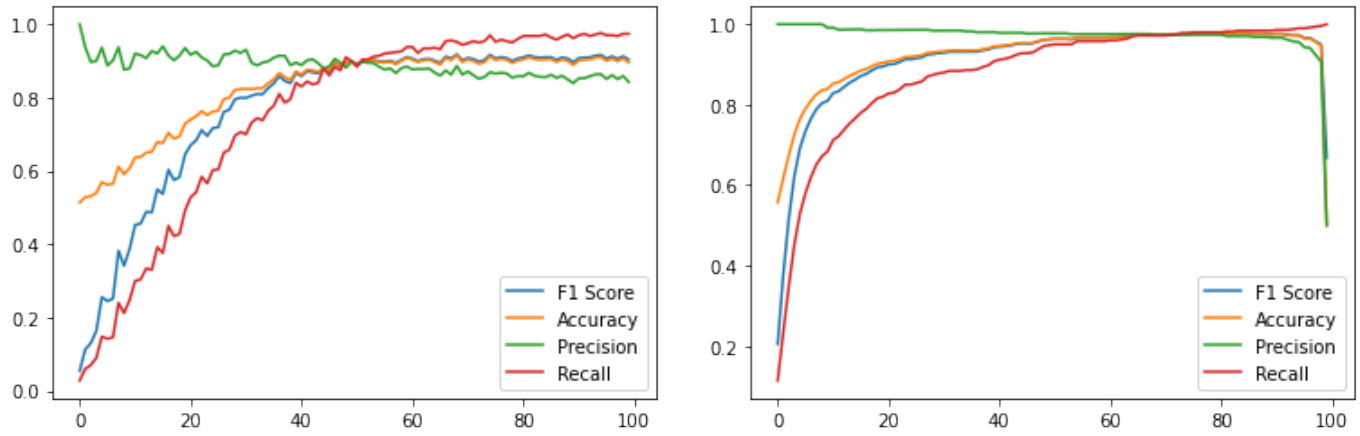
[1] Duong, Chi Thang, et al. "On node features for graph neural networks." arXiv preprint arXiv:1911.08795 (2019).

[2] Cai, Chen, and Yusu Wang. "A simple yet effective baseline for non-attributed graph classification." arXiv preprint arXiv:1811.03508 (2018).Cai, Chen, and Yusu Wang. "A simple yet effective baseline for non-attributed graph classification." arXiv preprint arXiv:1811.03508 (2018).

[3] Gómez-Adorno, H.; Sidorov, G.; Pinto, D.; Vilariño, D.; Gelbukh, A. Automatic Authorship Detection Using Textual Patterns Extracted from Integrated Syntactic Graphs. Sensors 2016, 16, 1374. https://doi.org/10.3390/s16091374

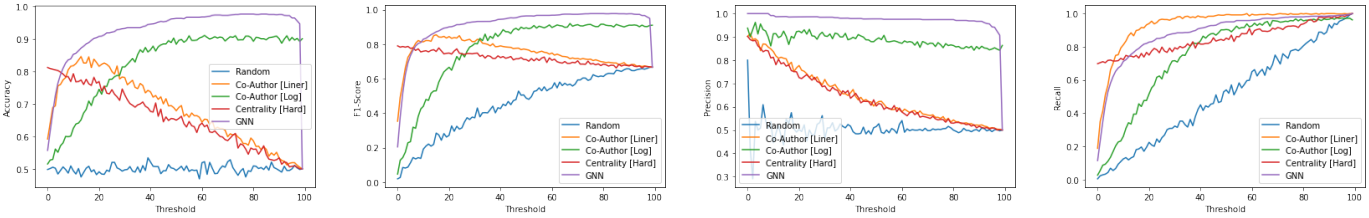
[4] Ravneet Kaur, Sarbjeet Singh, Harish Kumar, AuthCom: Authorship verification and compromised account detection in online social networks using AHP-TOPSIS embedded profiling based technique, Expert Systems with Applications, Volume 113, 2018, Pages 397-414, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2018.07.011.

[5] Castillo, Esteban, Cervantes, Ofelia, and Vilariño, Darnes. 'Authorship Verification Using a Graph Knowledge Discovery Approach'. 1 Jan. 2019 : 6075 – 6087.



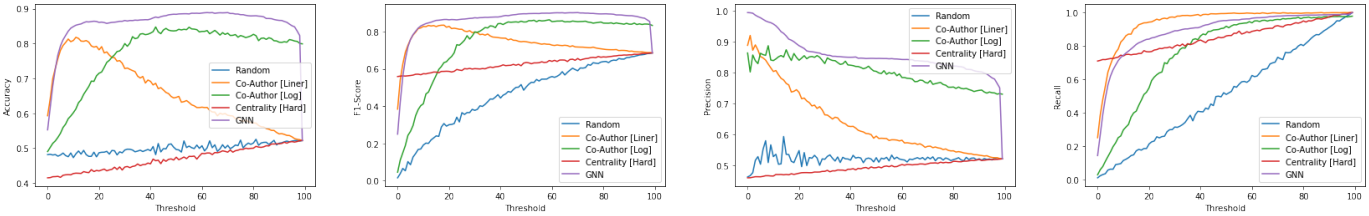
(a) (b)

Figure 5: Co-Author [Log] (Acc: 0.918, F1: 0.915) (a) and GASC (Acc: 0.988, F1: 0.988) (b)



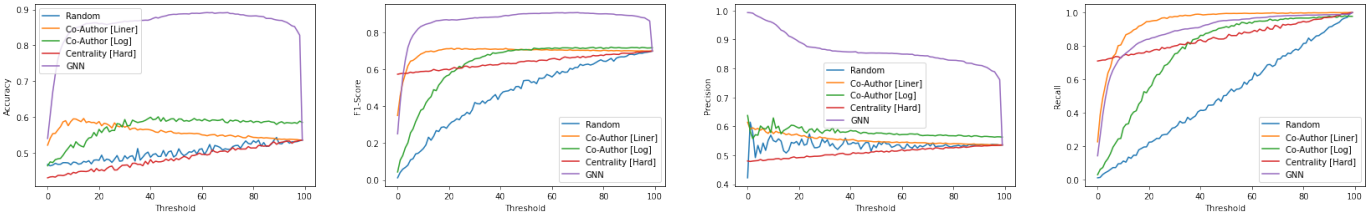
(a) (b) (c) (d)

Figure 6: Validation Set: Accuracy (a) F1-Score (b) Precision (c) Recall (d); GASC: 0.988, Base: 0.918



(a) (b) (c) (d)

Figure 7: Random Node: Accuracy (a) F1-Score (b) Precision (c) Recall (d); GASC: 0.850, Base: 0.828



(a) (b) (c) (d)

Figure 8: Challenging Cases: Accuracy (a) F1-Score (b) Precision (c) Recall (d); GASC: 0.872, Base: 0.598