EasyChair Preprint

№ 1297

Matching Anonymized Individuals with Errors for Service Systems

Wai Kin Victor Chan

July 17, 2019

Matching Anonymized Individuals with Errors for Service Systems

Wai Kin (Victor) Chan*

Environmental Science and New Energy Technology Engineering Laboratory
Tsinghua-Berkeley Shenzhen Institute, Tsinghua University
Shenzhen 518055, P.R. China

ABSTRACT

Data privacy is of great importance for the healthy development of service systems. Companies and governments that provide services to people often have big concerns in sharing their data. Because of that, data must be pre-processed (e.g., anonymized) before they can be shared. However, without identification, it is difficult to match data from different sources and thus the data cannot be used together. This paper investigates how the performance of two simple individual matching methods was affected by errors in the similarity scores between individuals. The first method is a greedy method (GM) that simply matches individuals based on the maximum similarity scores. The second method is an optimal assignment problem (AP), which maximizes the total similarity scores of the matched individuals. Consistent with the literature, we found that GM outperforms AP in most situations. However, we also discovered that AP could be better in fixing errors.

Key words: data matching, data correlation, service systems

1. INTRODUCTION

Value co-creation is one core principle of the development of service systems (Maglio et al. 2010). A number of studies have demonstrated the benefits of value co-creation and how it fosters the development of many business and social sectors (Spohrer and Maglio 2008; Hsu 2009; Vargo and Akaka 2012). Many innovations in service science require the use of user data to enable or create new services. However, privacy is a big concern nowadays due to the raising of big data. As people (e.g., customers) are using more and more services, their digital trails could be collected and used by unauthorized individuals or organizations. As such, it is important to protect privacy of customers while still allowing data to be used for service innovations.

One application is data matching. In data matching, two or more databases from different owners (e.g., service organizations) are matched based on individual identity. The combined dataset is a larger database that contains much richer information about each individual and the whole population in the dataset. This allows more accurate services to be provided to customers. Unfortunately, due to privacy issue owners of the databases cannot share their data unless at least identifications of individuals are encrypted or removed.

Privacy-preserving data matching (or record linkage) has been an active research area (Karakasidis and Verykios 2010; Christen 2012a; McCormack and Smyth 2017; Franke et al. 2018). One of the earliest work in data matching was done in (Fellegi and Sunter 1969). Data matching has a number of applications. (Kum et al. 2014) create an architecture for social genome databases by connecting social data from various sources. (Fu et al. 2011; Fu et al. 2014) use census data to match records at a household level. A graph-based approach was used to capture structural relationship between household members.

There are a number of data matching methods, based on blocking and indexing (Christen 2012b; Fisher et al. 2015). There are two basic types of protocols governing how data is exchanged between database owners: 1) two-party protocol and 3) three-party protocol. The three-party protocol requires a third party that is trusted by the database owners to do the matching of the data. The two-party protocol does not rely on the use of a third party. As such, sensitive personal information is removed or encrypted by using some encryption or encoding methods mutually agreed between the database owners. Many models have been studied (Karakasidis and Verykios 2010) and a number

* Corresponding author: Tel.: (86) 3688-1023; E-mail: chanw@sz.tsinghua.edu.cn

of advanced encryption approaches have also been proposed, such as secure hash algorithm, secure secret function evaluation, and message digest function (Kilian 1990; Schneier 1996).

The procedure of data matching involves the following steps in sequence: 1) data pre processing, 2) indexing, 3) comparison, 4) classification, 5) clerical review, and 6) evaluation. Data pre-processing is to prepare and clean the data. Indexing is to filter out unlikely matches to reduce the complexity of matching. The comparison step compares records of the matching databases and assigns each pair of records a similarity score. The classification step matches the records based on the similarity scores. If the matching method classifies records are possible matches, clerical review is needed to make decision. Finally, evaluation can be carried out to assess the matching quality.

This paper focuses on analyzing how matching quality is affected by errors, which could occur during the calculation of the similarity scores or simply due to data entry mistakes. To rule out influence from other factors, this paper will not apply any classification operations. In addition, we assume that the two matching databases are of the same size and correspond to the same *n* distinct individuals, that is, the matching is a one-to-one matching. Therefore, we assume that the *n*x*n* similarity scores have been obtained during, for example, the comparison step by using some similarity functions (Vatsalan et al. 2013). However, errors can occur either in record entries or during score calculation. Therefore, it is important to evaluate how such errors influence the matching quality. The assumption of two databases being equal size can be realized if the *n* individuals are fixed and the two databases contain data only for these *n* individuals.

As this paper only considers one-to-one matching, we focus on the two methods used in (Christen 2012a) and (Franke et al. 2018): 1) Greedy Method (GM) and 2) Assignment Problem (AP).

Although (Franke et al. 2018) also allow errors in their study, the present paper focuses on examining the impact of error rate to the matching quality. In addition, besides considering precision and recall, the present paper divides the errors into four types and explains why AP performed worse than GM. This shows counter-intuitively that optimization is not always better. We aim at extending the study in (Christen 2012a; Franke et al. 2018) by providing a detailed analysis in how errors affect matching quality.

Section 2 defines the problem and outline the two approaches, GM and AP. Section 3 presents the main results and performs an analysis on the experimental results. Section 4 draws a conclusion and offers several future works.

2. Problem Definition and Two Data Matching Methods: Greedy Method and Assignment Problem

The original problem of privacy-preserving data matching (or linkage) is to find a matching between two or more databases owned by different organizations in such a way that no sensitive information is shared across organizations.

This paper considers two post-matching methods: Greedy Method (GM) and optimal Assignment Problem (AP). As introduced in (Christen 2012a), the GM simply matches records based on the highest available similarity scores; that is, the pair of records with the highest similarity score are matched first and their associated links are removed. Then, the next pair of records with the highest similarity score (after removing links with the first pair) are matched and their associated links are removed. This process repeats until all records are matched. Note that because we do not employ any classification operations (i.e., no threshold is used), all records between the two databases are fully connected (i.e., no missing similarity score). Therefore, the GM will eventually obtain a complete match between databases.

AP is a special case of the network flow problem (Ahuja et al. 1993). Like many network flow problems, AP also satisfies the network property, meaning that linear relaxation solution is also integer solution. As such, existing network flow solution algorithm can solve this problem efficiently.

As explained in previous section, this paper assumes that the matching is conducted between two databases of equal size, and that the databases are deduplicated. Precisely, let D_A and D_B be the two databases to be matched and $|D_A| = |D_B| = n$, where n is the size (number of records) in each database. Let binary variable $\delta_{ij} = 1$ if Record i in D_A and Record j in D_B are a true match and $\delta_{ij} = 0$ otherwise. It is also assumed that there exists a true perfect matching between the two databases, that is, for each Record i in D_A, there exists one and only one Record j in D_B that corresponds to the same individual as Record i in D_A. However, this true perfect matching is unknown.

A premise of this study is that similarity scores between the records of the two databases are available. In other words, the comparison step will result in a set of similarity scores, $\mathbf{S} = \{s_{ij}, \forall i \in D_A, \forall j \in D_B\}$. In the literature, it is commonly assumed that the score is between 0 and 1. However, in this paper, we generalize it so it can take any real value (see next section).

Let binary variable $x_{ij} = 1$ if Record i in D_A is found to be a match with Record j in D_B and $x_{ij} = 0$ otherwise. The AP is to find $\mathbf{X} = \{x_{ij}, \forall i \in D_A, j \in D_B\}$ that satisfies the one-to-one matching constraints: $\sum_{\forall i \in D_A} x_{ij} = 1, \forall j \in D_B$ and $\sum_{\forall j \in D_B} x_{ij} = 1, \forall i \in D_A$. Formally, AP is defined as:

$$\max_{\mathbf{X}} f_1 = \sum_{\forall i,j} s_{ij} x_{ij}$$
$$s.t. \sum_{\forall i \in D_A} x_{ij} = 1, \quad \forall j \in D_B$$
$$\sum_{\forall j \in D_B} x_{ij} = 1, \quad \forall i \in D_A$$

3. EXPERIMENTS

In this short paper, we conduct an experiment on a network of 1000 individuals. Because no classification nor threshold is used, the size of the problem is *n*x*n*. This full size problem allows us to thoroughly examine the impact of error rate when it varies from 0.05 to 0.95. Also, because we use a synthetic network, we can designate the true matching individuals to evaluate the matching results. In particular, we set all diagonal elements to correspond to the same individuals.

We first generate the *n*x*n* similarity scores, $\mathbf{S} = \{s_{ij}, \forall i \in D_A, \forall j \in D_B\}$ according to a standard normal distribution. If these scores represent a real system, the similarity score of the pair of records pertaining to same individual should have the largest score. Hence, for each individual i, we swap s_{ii} with the maximum score among all s_{ij}, $j = 1,\ldots, n$. Therefore, the true matches are all diagonal elements in the *n*x*n* matrix.

The next step is to make some errors. Each score has a chance of p_e to be replaced by a new value following a standard normal distribution independent of the previous value. We vary p_e from 0.05 to 0.95 at an increment of 0.05. For each p_e, five replications were made. This error generation process creates two types of errors: 1) changing s_{ii} to be smaller than at least one other score, s_{ij} and 2) changing at least one s_{ij} to be larger than s_{ii}. It is also possible that s_{ii} is first reduced to be less than only one s_{ij} and then s_{ij} is changed to be smaller than the new s_{ii}. In such a case, no effective error is made.

After the error generation process, a synthetic set of *n*x*n* similarity scores with errors is obtained. If one uses this set of similarity scores to determine the matching (i.e., based on the maximum scores), the potential error in the matching is the fraction of non-maximum scores within the diagonal line.

Two standard accuracy measures are precision and recall. Precision is the percentage of correct matches within all "matches" (correct and wrong) found by the algorithm, that is, correct matches / (correct matches + wrong matches). Recall is the percentage of correct matches within all actual matches in the data, that is, correct matches / (correct matches + false actual matches). Because there exactly n individuals in this study, the denominator of both precision and recall are the same. That is, wrong matches (wrongly classified as matches) must equal false actual matches (true matches wrongly classified as un-matches). Therefore, in the case of one-to-one matching, precision is identical to recall.

Figure 1 presents the results of precision (same as recall). It is clear that GM out performs AP regardless of the error rate (the superiority is minor especially at high error rate). This finding is consistent with the results in (Franke et al. 2018). Also, as the error rate increases, the performance of both GM and AP degrade as expected.

Next, we extend the analysis by breaking down the matching results into four categories: 1) Match to Match (MM), 2) Match to Un-Match (MU), 3) Un-Match to Match (UM), and 4) Un-Match to Un-Match (UU). The diagonal element (i.e., s_{ii}) is called MM if no error occurs in s_{ii} (that is, the similarity score still represents the right match) (Match) and then after the matching algorithm (either GM or AP) it was still classified as the same person (i.e., $x_{ii} = 1$) (Match). The diagonal element is called MU if no error occurs (Match) but after the matching algorithm it was classified as different persons (i.e., $x_{ii} = 0$) (Un-Match). Similarly, UM means that an error occurs in s_{ii} (Un-Match) but was correctly classified (i.e., fixed) as the same person, and UU represents that an error occurs in s_{ii} (Un-Math) and the algorithm was unable to fix the error by classifying it as a match (Un-Match).

These four performance measures are shown in Figure 2.a-d. Figure 2.a and b are consistent with Figure 1.a in that GM obtains higher accuracy. Figure 2.c further supports this result. It shows that AP can turn more originally correct

matches into incorrect matches. The reason for this is that when AP tries to maximize the total similarity score, it could scarify a maximum similarity score to save several non-maximum scores so the total score is maximized.

Figure 2.d, however, reveals a situation where AP is better than GM: When examining UM, we found that AP, also due to its maximization objective, has a better chance to fix some erroneous un-matches.

Because of this special situation and to be fair, we define another performance measure: the "fixed to messed up ratio", which equals to UM/MU—the number of errors a method fixed divided by the number of errors it created. This measure for the two methods is shown in Figure 3. It shows that the "marginal benefit" (number of errors fixed per unit of errors created) of AP is higher than that of GM in most cases, in particular, when the error rate is higher than 10%. In privacy-preserving data matching, because identification information is removed, the error rate is likely to be high in computing the similarity scores. As such, AP may still be a better choice if the goal is to fix more errors.

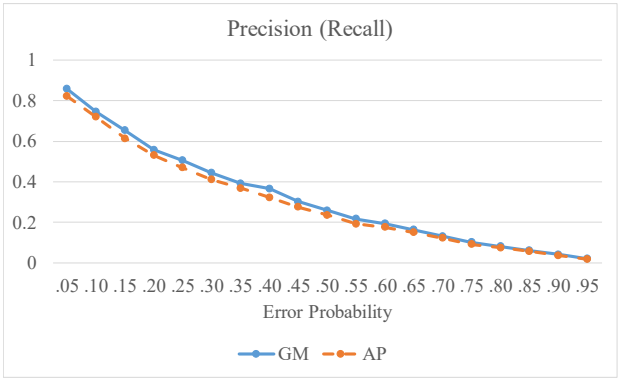
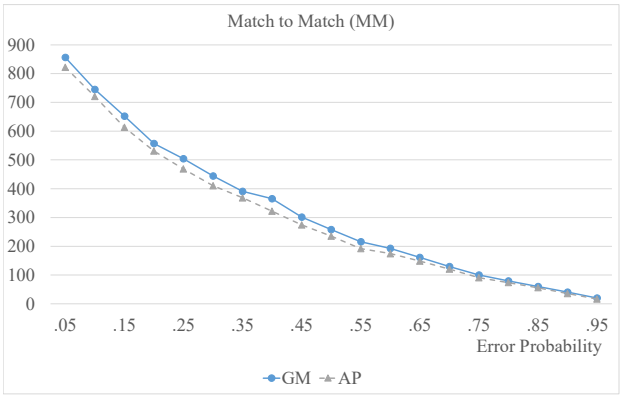
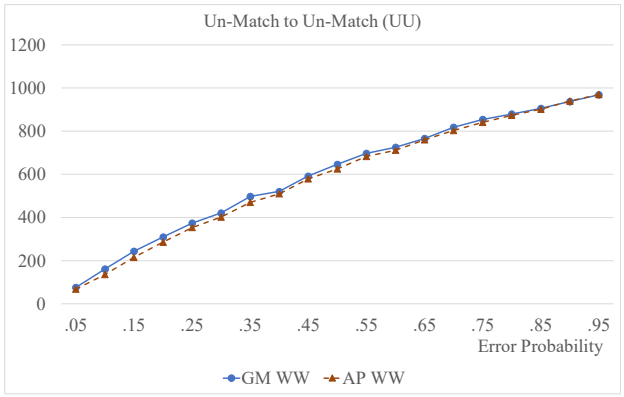


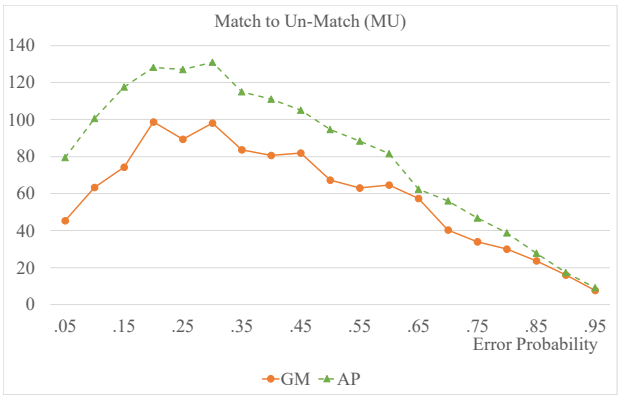
Figure 1: Performance (Precision or Recall) of GM and AP in changes of error rate

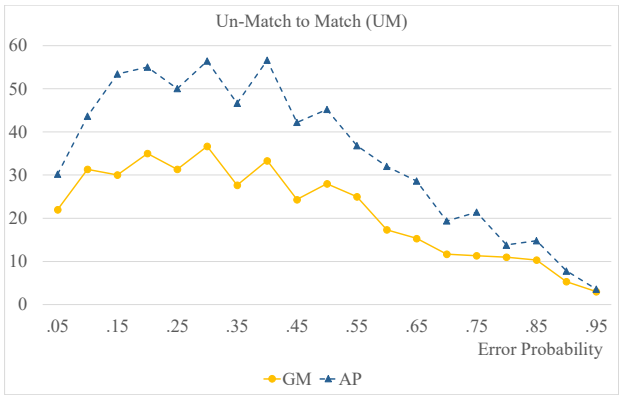


(a) Match to Match (MM)



(b) Un-Match to Un-Match (UU)





(c) Match to Un-Match (MU) (d) Un-Match to Match (UM)

Figure 2: Four performance measures.

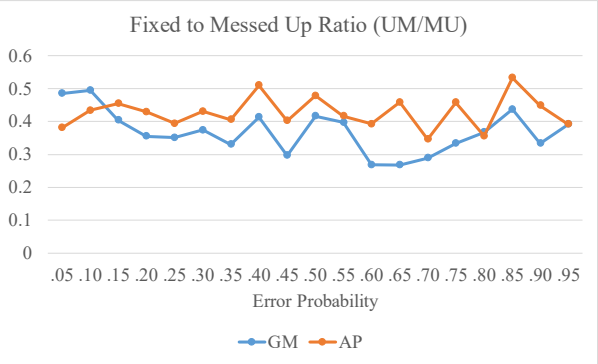


Figure 3: Fixed to Messed Up Ratio: UM/MU.

5. Conclusion and Future Work

Although studies in the literature found that the simple greedy method outperforms AP, we show that when a more thorough comparison is conducted, AP may still be better under the measure of error correction (i.e., UM). Due to space limitation, some of the results were not included. There are many possible directions for future work. First, real data should be used to examine the four performance measures. Second, larger networks should be used to see if the results change in the size of the network. Third, the four performance measures should be analyzed theoretically rather than just experimentally. Last but not least, the reason for AP to perform better GM under UM should be analyzed analytically.

Acknowledgements

This paper was partially funded by Shenzhen Municipal Development and Reform Commission, Shenzhen Environmental Science and New Energy Technology Engineering Laboratory, Grant Number: SDRC [2016]172.

References

Ahuja, R. K., T. L. Magnanti, and J. B. Orlin. 1993. *Network Flows: Theory, Algorithms, and Applications*: Prentice-Hall, Inc.

Christen, P. 2012a. *Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. edited by M. J. Carey, and S. Ceri. Berlin: Springer.

Christen, P. 2012b. "A Survey of Indexing Techniques for Scalable Record Linkage and Deduplication". *IEEE Transactions on Knowledge and Data Engineering* 24 (9):1537-1555.

Fellegi, I. P., and A. B. Sunter. 1969. "A Theory for Record Linkage". *Journal of the American Statistical Association* 64 (328):1183-1210.

Fisher, J., P. Christen, Q. Wang, and E. Rahm. 2015. "A Clustering-Based Framework to Control Block Sizes for Entity Resolution". In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, edited by 279-288. 2783396: ACM.

Franke, M., Z. Sehili, M. Gladbach, and E. Rahm. 2018. Post-Processing Methods for High Quality Privacy-Preserving Record Linkage, at Cham.

Fu, Z., P. Christen, and M. Boot. 2011. Automatic Cleaning and Linking of Historical Census Data Using Household Information. 2011 IEEE 11th International Conference on Data Mining Workshops, 11-11 Dec.

2011.
Fu, Z., P. Christen, and J. Zhou. 2014. A Graph Matching Method for Historical Census Household Linkage, at Cham.
Hsu, C. 2009. *Service Science: Design for Scaling and Transformation*. Singapore: World Scientific and Imperial College Press.
Karakasidis, A., and V. S. Verykios. 2010. "Advances in Privacy Preserving Record Linkage.
E-Activity and Innovative Technology, Advances in Applied Intelligence
Technologies Book Series, Igi Global, 2010.". In.
Kilian, J. 1990. *Uses of Randomness in Algorithms and Protocols*. Cambridge, Massachusetts: MIT Press.
Kum, H., A. Krishnamurthy, A. Machanavajjhala, and S. C. Ahalt. 2014. "Social Genome: Putting Big Data to Work for Population Informatics". *Computer* 47 (1):56-63.
Maglio, P., C. Kieliszewski, and J. Spohrer. 2010. *Handbook of Service Science*. New York, NY: Springer.
McCormack, K., and M. Smyth. 2017. "Privacy Protection for Big Data Linking Using the Identity Correlation Approach". *统计科学与应用: 英文版* (3):81-90.
Schneier, B. 1996. *Applied Cryptography: Protocols, Algorithms, and Source Code in C*. 2nd ed. New York: John Wiley & Sons, Inc.
Spohrer, J., and P. P. Maglio. 2008. "The Emergence of Service Science: Toward Systematic Service Innovations to Accelerate Co-Creation of Value". *Production and Operations Management* 17 (3):238-246.
Vargo, S. L., and M. A. Akaka. 2012. "Value Cocreation and Service Systems (Re)Formation: A Service Ecosystems View". *Serv. Sci.* 4 (3):207-217.
Vatsalan, D., P. Christen, and V. S. Verykios. 2013. "A Taxonomy of Privacy-Preserving Record Linkage Techniques". *Information Systems* 38 (6):946-969.