

8. REFERENCES

- [1] M. Bayati, M. Gerritsen, D. F. Gleich, A. Saberi, and Y. Wang. Algorithms for Large, Sparse Network Alignment Problems. *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pages 705–710, 2009.
- [2] A. Cecaj, M. Mamei, and N. Biccocchi. Re-identification of anonymized CDR datasets using social network data. In *Pervasive Computing and Communications Workshops (PERCOM Workshops), 2014 IEEE International Conference on*, pages 237–242. IEEE, 2014.
- [3] A. Cecaj, M. Mamei, and F. Zambonelli. Re-identification and information fusion between anonymized CDR and social network data. *Journal of Ambient Intelligence and Humanized Computing*, 7(1):1–14, 2015.
- [4] E. Cho, S. A. Myers, and J. Leskovec. Friendship and mobility: user movement in location-based social networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1082–1090. ACM Request Permissions, 2011.
- [5] P. Christen. *Data Matching, Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2012.
- [6] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. M. Kleinberg. Inferring social ties from geographic coincidences. *Proceedings of the National Academy of Sciences*, 107(52):22436–22441, 2010.
- [7] Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. Unique in the Crowd: The privacy bounds of human mobility. *Scientific Reports*, 3, 2013.
- [8] Y.-A. de Montjoye, L. Radaelli, V. K. Singh, and A. S. Pentland. Unique in the shopping mall: on the reidentifiability of credit card metadata. *Science*, 347(6221):536–539, 2015.
- [9] O. Goga, H. Lei, S. Parthasarathi, and G. Friedland. Exploiting innocuous activity for correlating users across sites. In *WWW '13: Proceedings of the 22nd international conference on World Wide Web*, pages 447–458, 2013.
- [10] O. Goga, P. Loiseau, R. Sommer, R. Teixeira, and K. Gummadi. On the Reliability of Profile Matching Across Large Online Social Networks. In *KDD '15: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1799–1808. ACM Request Permissions, 2015.
- [11] S. Ji, W. Li, M. Srivatsa, J. S. He, and R. Beyah. Structure Based Data De-Anonymization of Social Networks and Mobility Traces. In *ISC Proceedings of the 17th International Information Security Conference*, pages 237–254. Springer International Publishing, 2014.
- [12] E. Kazemi, S. H. Hassani, and M. Grossglauser. Growing a graph matching from a handful of seeds. *Proceedings of the VLDB Endowment*, 8(10):1010–1021, 2015.
- [13] N. Korula and S. Lattanzi. An efficient reconciliation algorithm for social networks. *Proceedings of VLDB*, 7(5):377–388, 2014.
- [14] D. Koutra, H. Tong, and D. Lubensky. BIG-ALIGN: Fast Bipartite Graph Alignment. In *Data Mining (ICDM), 2013 IEEE 13th International Conference on*, pages 389–398, 2013.
- [15] A. Narayanan and V. Shmatikov. Robust De-anonymization of Large Sparse Datasets. *Security and Privacy, 2008. SP 2008. IEEE Symposium on*, pages 111–125, 2008.
- [16] A. Narayanan and V. Shmatikov. De-anonymizing Social Networks. *Security and Privacy, 2009 30th IEEE Symposium on*, pages 173–187, 2009.
- [17] P. Pedarsani and M. Grossglauser. On the privacy of anonymized networks. In *KDD '11: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1235–1243. ACM Request Permissions, 2011.
- [18] C. J. Riederer, S. Zimmeck, C. Phanord, A. Chaintreau, and S. M. Bellovin. I don't have a photograph, but you can have my footprints.: Revealing the Demographics of Location Data. In *COSN '15: Proceedings of the third ACM conference on Online social networks*, pages 185–195. ACM, 2015.
- [19] L. Rossi and M. Musolesi. It's the Way you Check-in: Identifying Users in Location-Based Social Networks. *COSN '14: Proceedings of the 2nd ACM conference on Online social networks*, pages 215–226, 2014.
- [20] M. Srivatsa and M. Hicks. De-anonymizing Mobility Traces: Using Social Networks as a Side-Channel. *CCS '12: Proceedings of the 2012 ACM conference on Computer and communications security*, pages 628–637, 2012.
- [21] L. Sweeney. k-anonymity: a model for protecting privacy. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 10(5):557–570, 2002.
- [22] J. Unnikrishnan. Asymptotically Optimal Matching of Multiple Sequences to Source Distributions and Training Sequences. *Information Theory*, 61(1):452–468, 2015.
- [23] J. Unnikrishnan and F. M. Naini. De-anonymizing private data by matching statistics. In *Communication, Control, and Computing (Allerton), 2013 51st Annual Allerton Conference on*, pages 1616–1623. IEEE, 2013.
- [24] L. Yartseva and M. Grossglauser. On the performance of percolation graph matching. In *COSN '15: Proceedings of the third ACM conference on Online social networks*, pages 119–130. ACM Request Permissions, 2013.
- [25] H. Zang and J. Bolot. Anonymization of location data does not work: a large-scale measurement study. In *MobiCom '11: Proceedings of the 17th annual international conference on Mobile computing and networking*, pages 145–156. ACM Request Permissions, 2011.
- [26] J. Zhang, X. Kong, and P. S. Yu. Transferring heterogeneous links across location-based social networks. In *WSDM '14: Proceedings of the 7th ACM international conference on Web search and data*

mining, pages 303–312. ACM Request Permissions, 2014.

- [27] Y. Zhong, N. J. Yuan, W. Zhong, F. Zhang, and X. Xie. You Are Where You Go. In *WSDM '15: Proceedings of the 8th ACM international conference on Web search and data mining*, pages 295–304. ACM Press, 2015.

9. APPENDIX

9.1 Proof of Theorem 1

We first show that each of the 2 factors in the denominator of $\phi(a_1, a_2)$ can be replaced by the corresponding truncated sum while affecting its value by at most $1 + 1/C^2$. Since the numerator is decreased by truncation, this establishes the upper bound on $\phi'(a_1, a_2)$. We then show that for the numerator of $\phi(a_1, a_2)$, the difference between the infinite sum and its truncated version is at most $1/C$ times the first term in this sum. Since the denominator is decreased by truncation, this establishes the lower bound on ϕ' .

To obtain the upper bound, we first consider the factor $\sum_{k=a_1}^{\infty} \frac{\lambda^k}{k!} \binom{k}{a_1} (1-p_1)^{k-a_1}$ in the denominator. Expanding the binomial coefficient and pulling common terms outside the summation, this factor can be written as:

$$\frac{\lambda^{a_1}}{a_1!} \sum_{k \geq a_1} \frac{\lambda^{k-a_1} (1-p_1)^{k-a_1}}{(k-a_1)!} = \frac{\lambda^{a_1}}{a_1!} \sum_{k \geq 0} \frac{\lambda^k (1-p_1)^k}{k!}$$

Note that first term in this revised sum evaluates to 1, the term of index $\ln C$ evaluates to $\lambda^{\ln C} (1-p_1)^{\ln C} / (\ln C)! \ll \frac{1}{C^2}$, and the sum of all terms from $\ln C$ onward are at most $\frac{\lambda^{\ln C} (1-p_1)^{\ln C} / (\ln C)!}{(1-\lambda)}$ (upper bounding the infinite sum with a geometric series). Since $\lambda < 1/2$, we conclude that the sum of all terms from index $\ln C$ onward are less than $1/C^2$ times the first term.

The truncated sum for the second factor in the denominator can be bounded identically, giving us the desired upper bound on $\phi'(a_1, a_2)$.

It remains only to establish the lower bound by bounding the truncated numerator. We assume without loss of generality that $a_1 \geq a_2$. Expanding the binomial coefficients in the definition of the numerator of $\phi(a_1, a_2)$ and pulling common terms outside the summation, we can rewrite the numerator as:

$$\frac{\lambda_1^{a_1} (1-p_2)^{(a_1-a_2)}}{a_1! a_2!} \sum_{k \geq a_1} \frac{\lambda^{k-a_1} ((1-p_1)(1-p_2))^{k-a_1} \cdot k!}{(k-a_1)! (k-a_2)!}$$

The first term inside the revised sum is simply $a_1! / (a_1 - a_2)! > 1$. Let i denote the final index in the truncated sum, $a_1 + \max\{\ln C, 2a_1\}$. The i th term is upper bounded by $\lambda^{i-a_1} \cdot \frac{i!}{(i-a_1)! (i-a_2)!}$. If $a_1 \geq 4$, then since $i \geq 3a_1$, it is easy to see that $\frac{i!}{(i-a_1)!^2} < 1/2$. If $a_1 \leq 4$, then since $i - a_1 \geq \ln C \geq 7$, we can note that $\frac{i!}{(i-a_1)!^2} < 1/2$. As $\lambda < 1/2$ and $i > a_1 + \ln C$, the i th term is less than $1/C \cdot 1/2$. Again upper bounding the infinite sum with a geometric series, the sum of all terms from index i onward is less than the i th term divided by $(1-\lambda)$, and hence $< 1/C$. Therefore, the sum of all terms from the i th term onward is less than $1/C$ times the first term, completing the proof.

9.2 Proof of Lemma 2

Recall that in Lemma 2, we proved that $E[\text{Score}(u, v, \ell, t)] \leq 0$ for any pair of users u, v such that $v \neq \sigma_I(u)$. For $v = \sigma_I(u)$, we showed that the expected score is lower bounded by:

$$\begin{aligned} & X(0, 0) \ln \frac{X(0, 0)}{Y(0, 0)} + (1 - X(0, 0)) \ln \frac{(1 - X(0, 0))}{(1 - Y(0, 0))} \\ &= X(0, 0) \ln \frac{X(0, 0)}{Y(0, 0)} - (1 - X(0, 0)) \ln \frac{(1 - Y(0, 0))}{(1 - X(0, 0))} \\ &\geq (1 - \lambda(p_1 + p_2 - p_1 p_2)) \lambda p_1 p_2 - \\ &\quad \lambda(p_1 + p_2 - p_1 p_2) \ln \frac{(1 - e^{-\lambda(p_1 + p_2)})}{(1 - e^{-\lambda(p_1 + p_2 - p_1 p_2)})} \end{aligned}$$

To prove that this expression is lower bounded by $(\lambda p_1 p_2)^2 K$, it suffices to prove that:

$$\begin{aligned} & (1 - \lambda(p_1 + p_2 - p_1 p_2)) \lambda p_1 p_2 - \\ & \lambda(p_1 + p_2 - p_1 p_2) \ln \frac{(1 - e^{-\lambda(p_1 + p_2)})}{(1 - e^{-\lambda(p_1 + p_2 - p_1 p_2)})} \\ & \geq (\lambda p_1 p_2)^2 K \end{aligned}$$

or equivalently:

$$\begin{aligned} & (1 - \lambda(p_1 + p_2 - p_1 p_2)) p_1 p_2 - \lambda(p_1 p_2)^2 K \\ & - (p_1 + p_2 - p_1 p_2) \ln \frac{(1 - e^{-\lambda(p_1 + p_2)})}{(1 - e^{-\lambda(p_1 + p_2 - p_1 p_2)})} \geq 0 \quad (2) \end{aligned}$$

We can simplify the final factor in this inequality as follows:

$$\begin{aligned} & \ln \frac{(1 - e^{-\lambda(p_1 + p_2)})}{(1 - e^{-\lambda(p_1 + p_2 - p_1 p_2)})} = \ln e^{-\lambda(p_1 p_2)} \frac{(e^{\lambda(p_1 + p_2)} - 1)}{(e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1)} \\ &= \left(\ln \frac{(e^{\lambda(p_1 + p_2)} - 1)}{(e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1)} \right) - \lambda p_1 p_2 \end{aligned}$$

where the first equality came from multiplying the numerator and denominator by $e^{\lambda(p_1 + p_2 - p_1 p_2)}$.

Substituting into Inequality (2), our lemma reduces to:

$$\begin{aligned} & (1 - \lambda(p_1 + p_2 - p_1 p_2)) p_1 p_2 - \lambda(p_1 p_2)^2 K \\ & (p_1 + p_2 - p_1 p_2) \left(\ln \frac{(e^{\lambda(p_1 + p_2)} - 1)}{(e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1)} - \lambda p_1 p_2 \right) \geq 0 \end{aligned}$$

or, equivalently:

$$\begin{aligned} & p_1 p_2 (1 - \lambda(p_1 p_2) K) - \\ & (p_1 + p_2 - p_1 p_2) \ln \frac{(e^{\lambda(p_1 + p_2)} - 1)}{(e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1)} \geq 0 \quad (3) \end{aligned}$$

This is hard to simplify directly, so we introduce the following upper bound:

$$\lambda p_1 p_2 = \ln \frac{1}{e^{-\lambda p_1 p_2}} = \ln \frac{e^{\lambda(p_1 + p_2)}}{e^{\lambda(p_1 + p_2 - p_1 p_2)}} \leq \ln \frac{e^{\lambda(p_1 + p_2)} - 1}{e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1}$$

Using Z to represent the quantity $\ln \frac{e^{\lambda(p_1 + p_2)} - 1}{e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1}$ and substituting the new inequality in Inequality (3), we are try-

ing to prove:

$$\begin{aligned}
& p_1 p_2 (1 - ZK) - (p_1 + p_2 - p_1 p_2) Z \geq 0 \\
& \Leftrightarrow p_1 p_2 \geq (p_1 + p_2 - p_1 p_2 (1 - K)) Z \\
& \Leftrightarrow \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2 (1 - K)} \geq Z \\
& \Leftrightarrow e^{\frac{p_1 p_2}{p_1 + p_2 - p_1 p_2 (1 - K)}} \geq \frac{e^{\lambda(p_1 + p_2)} - 1}{e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1}
\end{aligned}$$

Now to conclude the proof we use two inequalities that follows from the Taylor expansions. In particular we have:

$$e^x \geq 1 + x + \frac{1}{2}x^2$$

and for $x \in o(1)$:

$$e^x \leq 1 + x + x^2$$

Now by assuming that $\lambda \in o(1)$ and by fixing $K = \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2$ we get:

$$\begin{aligned}
& e^{\frac{p_1 p_2}{p_1 + p_2 - p_1 p_2 (1 - K)}} \geq \frac{e^{\lambda(p_1 + p_2)} - 1}{e^{\lambda(p_1 + p_2 - p_1 p_2)} - 1} \\
\Leftrightarrow & 1 + \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2} + \\
& \frac{p_1^2 p_2^2}{2(p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2)^2} \geq \\
& \frac{\lambda(p_1 + p_2) + \lambda^2(p_1 + p_2)^2}{\lambda(p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2)} \\
\Leftrightarrow & 1 + \frac{p_1 p_2}{p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2} + \\
& \frac{p_1^2 p_2^2}{2(p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2)^2} \geq \\
& 1 + \frac{p_1 p_2 + \lambda(p_1 + p_2)^2}{p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2} \\
\Leftrightarrow & \frac{\frac{1}{2}p_1^2 p_2^2}{p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2} \geq \lambda(p_1 + p_2)^2
\end{aligned}$$

Now by fixing $\lambda < \frac{1}{8} \frac{p_1^2 p_2^2}{(p_1 + p_2)^2}$ we get:

$$\begin{aligned}
& \frac{\frac{1}{2}p_1^2 p_2^2}{p_1 + p_2 - p_1 p_2 + \frac{1}{2}\lambda(p_1 + p_2 - p_1 p_2)^2} \geq \lambda(p_1 + p_2)^2 \\
\Leftrightarrow & \frac{\frac{1}{2}p_1^2 p_2^2}{p_1 + p_2 - p_1 p_2 + \frac{1}{16}p_1^2 p_2^2} \geq \frac{1}{8}p_1^2 p_2^2 \\
\Leftrightarrow & \frac{1}{4}p_1^2 p_2^2 \geq \frac{1}{8}p_1^2 p_2^2
\end{aligned}$$

So the claim follows.