Proceedings of the 3rd International Conference on Statistics: Theory and Applications (ICSTA'21)
~~Prague, Czech Republic~~ Virtual Conference – July, 2021
Paper No. 136
DOI: 10.11159/icsta21.136

The Impact of Entity Resolution on Observed Social Network Structure

Abby Smith[1]
[1]Northwestern University
2006 Sheridan Rd., Evanston, Illinois 60626
als1@u.northwestern.edu

Extended Abstract

Deduplication, also referred to as "entity resolution", is a common and crucial pre-processing step in the construction of social networks [1]. Citation network studies have indicated that false “splitting” and “lumping” of nodes can have dramatic downstream network impacts, and choices in deduplication methods are important for network analysis [2] [3]. Traditional deduplication methods compare the attributes (such as name and age) of potential matching pairs to estimate a match probability for a pair. Fellegi and Sunter (1969) [4] introduced an optimal decision threshold where above a certain matching score, pairs are declared a match, and below that threshold, pairs are considered a non-match. Recently research has used clustering techniques for entity resolution, where each cluster represents a unique underlying entity. Collective clustering techniques, pioneered by Bhattacharya and Getoor (2007) [5], relax unrealistic assumptions made by earlier probabilistic entity resolution techniques and allow matching decisions to be made dependent on each other. In social network datasets, we can also use relational information (e.g., a person’s network ties) in deduplication as further evidence for matching status of pair.

Entity resolution is inherently an imperfect process and is an outcome of existing measurement error, particularly when there is a lack of a manually-reviewed, "ground-truth" dataset to rely on for parameter tuning in a chosen technique [6]. I focus on two tuning parameters: the match decision threshold (t) in Felligi-Sunter, and the alpha trade-off parameter between attributional and relational similarity in Bhattacarya-Getoor. My work is focused on methods for evaluating entity resolution in a network setting, measuring the sensitivity of entity resolution results to choices in tuning parameters (alpha and t), and the downstream impacts these parameter choices can have on network metrics and topologies such as degree, closeness, and connectivity. I apply the evaluation methods to two real-world ego-centric network studies, (i) Care2Hope, a respondent-driven sample of rural people who use drugs (PWUD) in Appalachian Kentucky [1], and (ii) RADAR, a longitudinal network study of young men in Chicago who have sex with men. I consider evaluation scenarios in both the presence [7] and absence [8] of “ground truth” data . I discuss implications these findings could have for drug use and HIV policy, and make reporting recommendations for network analysts.

References

[1] Young, A. M., Rudolph, A. E., Su, A. E., King, L., Jent, S., & Havens, J. R. (2016). Accuracy of name and age data provided about network members in a social network study of people who use drugs: Implications for constructing sociometric networks. Annals of Epidemiology, 26(11), 802–809. https://doi.org/10.1016/j.annepidem.2016.09.010

[2] Fegley, B. D., & Torvik, V. I. (2013). Has Large-Scale Named-Entity Network Analysis Been Resting on a Flawed Assumption? PLoS ONE, 8(7), e70299. https://doi.org/10.1371/journal.pone.0070299

[3] Diesner, J., Evans, C. S., & Kim, J. (n.d.). Impact of Entity Disambiguation Errors on Social Network Properties. 10.

[4] Fellegi, I. P., & Sunter, A. B. (1969). A Theory for Record Linkage. Journal of the American Statistical Association, 64(328), 1183–1210. https://doi.org/10.1080/01621459.1969.10501049

[5] Bhattacharya, I., & Getoor, L. (2007). Collective entity resolution in relational data. ACM Transactions on Knowledge Discovery from Data, 1(1), 5-es. https://doi.org/10.1145/1217299.1217304

[6] Wang, D. J., Shi, X., McFarland, D. A., & Leskovec, J. (2012). Measurement error in network data: A re-classification. Social Networks, 34(4), 396–409. https://doi.org/10.1016/j.socnet.2012.01.003

[7] Harron, K. L., Doidge, J. C., Knight, H. E., Gilbert, R. E., Goldstein, H., Cromwell, D. A., & van der Meulen, J. H. (2017). A guide to evaluating linkage quality for the analysis of linked data. International Journal of Epidemiology, 46(5), 1699–1710. https://doi.org/10.1093/ije/dyx177

[7] Fisher, J., & Wang, Q. (2015). Unsupervised Measuring of Entity Resolution Consistency. 2015 IEEE International Conference on Data Mining Workshop (ICDMW), 218–221. https://doi.org/10.1109/ICDMW.2015.162