Sadowsky et al evaluate conjunction-based and hashing methods to block candidate pairs of death records from the Syrian civil war. The paper, "Blocking Methods Applied to Casualty Records from the Syrian Conflict" addresses the question of how to partition a collection of records into clusters that are likely to be matches without splitting matches across blocks. The impetus for this step is to reduce the need to conduct pairwise comparisons among all records in the data, a problem which quickly becomes intractable as the number of records grows. The paper outlines the results of applying locality sensitive hashing (LSH) to a dataset of names, dates of death, and locality of death for enumerated victims of the Syrian conflict. The task is particularly difficult as the ground-truth for the data is both unknown and unknowable. The three main approaches presented in the paper—transitive locality sensitive hashing (TLSH), k-means locality sensitive hashing (KLSH), and densified one permutation hashing (DOPH)—exemplify the diverse literature that record linkage problems intersect. The are drawn, respectively, from the community discovery, the information retrieval, and the near neighbor search[1] areas of research.

One pervasive issue in the Syrian data is that the training data on which they test is developed by hand-matching records. Interestingly, the criteria for the matchers is described as dividing records into "matches," which are the same individual, and a second set of records that have "no possibility of matching" [2, 6]. The criteria suggests that the test data has an extremely conservative threshold for a non-match: as described, one might imagine a large number of records that are neither unambiguous matches nor so different so as to be categorically unable to be the same entity.[2] Although HRDAG-affiliates have noted elsewhere that experienced matchers tend to become more conservative in linking records, the finality of declaring a non-match in the protocol could easily drive overmatching.

In terms of data-generation processes that might engender a baseline closer to the ground truth, I wonder if drone strike records would provide an illuminating set of data to test the three hash-based blocking approaches outlined in the paper. In such a case, once you can identify a person who was in the location, you may be able to use that information along with other contextual records about who else was at the same place at the same time (such as via family groupings or an event invitation record) would provide an additional information about unique individuals who should be identified. Essentially, the problem would be similar to a network sampling problem, in which network ties are co-location at the time of the strike.

[1] If that is a distinct literature

[2] Indeed, Fellegi and Sunter identify a third category, possible matches [1].

Adventures in LSH

In order to try to get a more intuitive sense of LSH, I implemented a simple LSH on a corpus of 5,000 news stories of events in Yemen. The articles date from January 17, 2000 through November 17, 2003 and were gathered for the ICEWS[3] The ICEWS event database codes event actions, but also includes the (English or English-translated) text of the news article on which the entry was based. Additionally, I carried out a similar analysis on 5,000 news articles from Bangladesh, although there were significantly fewer duplicates in this data.[4] The data is interesting from a record linkage perspective because the ICEWS data suffers from two duplication problems, one overt and the other more subtle. The overt problem is that news agencies often reprint stories, creating ICEWS entries that are virtually identical. The more subtle issue is that a single event between a particular pair of actors may generate many articles.[5]

For the LSH implementation, I leaned heavily on a tutorial for Locality Sensitive Hashing in R [6]. As I did not implement the Densified One Permutation Hashing (DPOH) algorithm, I wasn't able to benefit from the efficiency that DPOH brings. Thus, I limited the comparisons to the first 5,000 entries in the data.[7] Because I used only a small portion of the data, I also carried out pairwise Jaccard comparisons for each of the entries in the Bangladesh and Yemen. These comparisons are identified in black lines as the "truth" in the plots shown in Figure 1 The Yemen results presented below represent parsing articles into i) 2-word shingle and i) 3-word shingles, and 10 permutations of the identity matrix.[8] I did not do any additional text processing, such as removal of stop words, spaces and punctuation, or stray code markup.

One immediate issue that the exercise raised is the difficulty in establishing a threshold of interest for Jaccard similarity measures. Articles with a Jaccard similarity of 1 are clearly nearly identical; but they are relatively easy to match within the corpus. However, for both the Bangladesh and Yemen data, the results are concentrated at Jaccard Similarity measures near 0. Yet, prior knowledge of the data indicates that the data should be expected to have endemic duplication issues.

A second observation is that it would be extremely useful to incorporate structural features into the identity matrix. In particular, if one could incorporate time, it would be interesting to see how the use of certain phrases (or particular shingles) fluctuate across the time span of the data. The fact that the stories are sorted by date and numbered sequentially does give an indirect perspective on time, but the measure is very fragile.

[3]ICEWS, or the Integrated Crisis Warning System, is an event data set that consists of machine-coded news reports of interactions among political actors.

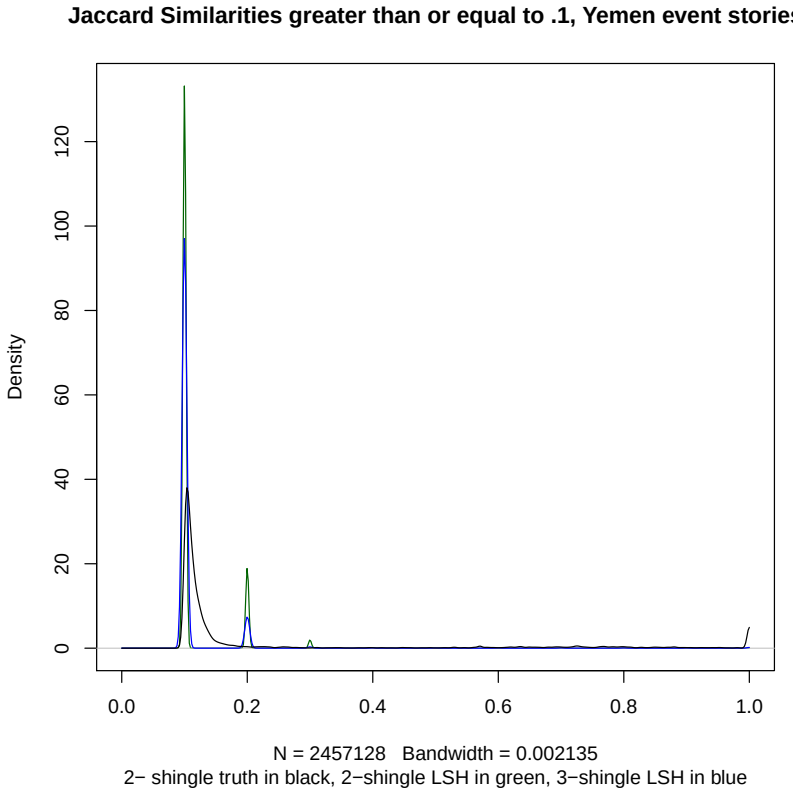
[4]The rate of duplicates is much lower because I used story data that I partially hand-cleaned two years ago.

[5]Consider, for example, the number of articles reporting on each controversial statement in the current election.

[6]Found at `http://dsnotes.com/articles/locality-sensitive-hashing-in-r-part-1`

[7]And, in the process, developed greater respect for runtime consequences.

[8]See above comment about renewed respect for runtime implications of hashing algorithms



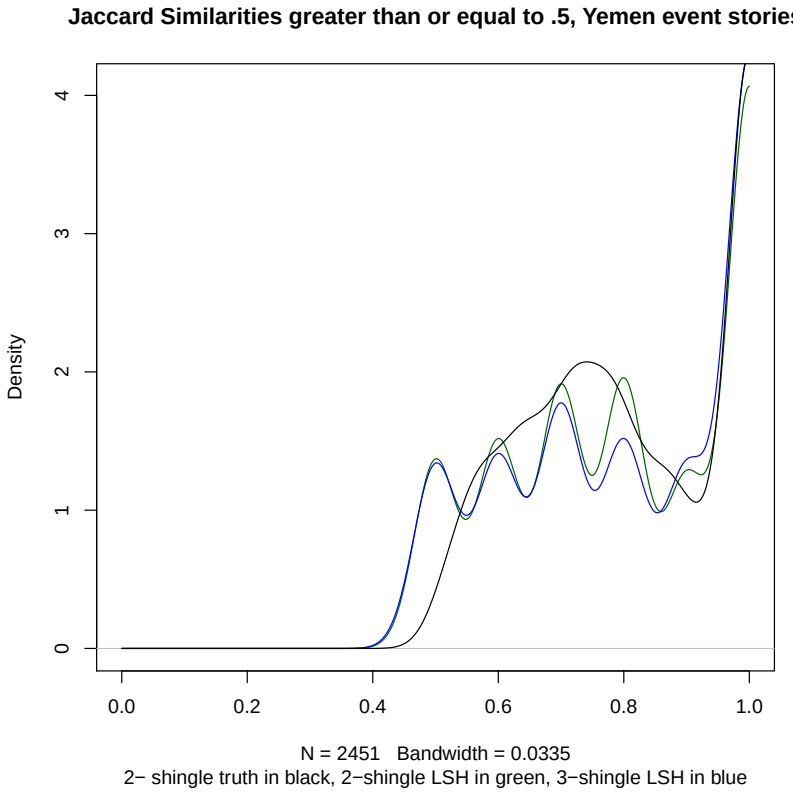


Table 1: Densities for Jaccard Similarities of Yemeni news stories

From a record-linkage perspective, incorporating additional information into the identity matrix may influence how to interpret Jaccard scores. Records that are relatively close matches and which were published at similar times are more likely to refer to the same event than are records with a similar Jaccard score but which were published months, or years, apart.

Thirdly, the process also highlighted data-presentation challenges when considering how to represent pairs with similar hashes. Graphs and plots make information more accessible, and it would be interesting to see how the data interrelate and cluster.

Finally, one question that I still have is what information is useful to extract from the hashing process. For example, at one point I plotted the distribution of Jaccard Scores that represented 10% quintiles for each document as compared to each other document. The resulting plot was interesting—particularly to compare patterns across the two different sets of news stories. However, I'm not sure whether the results were meaningful, in the sense of giving insight into underlying patterns of the data or outcomes of the LSH.

References

[1] Ivan P Fellegi and Alan B Sunter. A theory for record linkage. *Journal of the American Statistical Association*, 64(328):1183–1210, 1969.

[2] Peter Sadosky, Anshumali Shrivastava, Megan Price, and Rebecca C Steorts. Blocking methods applied to casualty records from the syrian conflict. *arXiv preprint arXiv:1510.07714*, 2015.