D-Dupe: An Interactive Tool for Entity Resolution in Social Networks

Mustafa Bilgic, Louis Licamele, Lise Getoor, and Ben Shneiderman

Computer Science Department, University of Maryland,
College Park, MD 20742, USA
{mbilgic, licamele, getoor, ben}@cs.umd.edu

Abstract. Graphs describing real world data often contain duplicate entries for names, cities, or other entities. This paper presents D-Dupe, an interactive visualization tool designed to help users to discover and resolve duplicate nodes in a social network. Users can resolve the ambiguity by merging nodes, or by specifying that the nodes are in fact distinct. The entity resolution process is iterative; as pairs of nodes are merged, additional duplicates may become apparent.

1 Introduction

The typical assumption in network visualization is that the underlying data is clean and the nodes refer to distinct entities while edges represent unique relationships. However, this presumption is rarely true. Networks are extracted from databases which may contain errors and inconsistencies. As data collection increases, and the databases themselves are being extended through automatic extraction techniques, or through the combination of multiple sources, the duplicate entries become more common place. This is especially true when the duplicate entries are not identical but slight variations of one another like abbreviations.

Duplicates may lead to inappropriate conclusions when the underlying data is visualized. Consider the example of a citation graph in which nodes correspond to authors and the node sizes are drawn in proportional to the number of publications of each author. If an author's name had multiple spellings and each of them were treated as distinct, the true entity will be represented not as one large node but as many unrelated small nodes. With such data (and representation), our conclusions about which author is most prolific will be incorrect.

In many cases, using the underlying structure of the network helps in resolving duplicates. For instance, suppose a bibliographic dataset that consists of four author references: "James Smith," "John Smith," "J. Smith," and "Mary Ann." We are interested in determining if "J. Smith" refers to "James Smith," to "John Smith" or it is a distinct author. Given no other information, we have no choice but to guess. But, if we know that "Mary Ann" collaborates with "John Smith" and "J. Smith" we are more likely to believe that "J. Smith" and "John Smith" refer to the same author. D-Dupe utilizes this idea by making such underlying structures apparent to users.

P. Healy and N.S. Nikolov (Eds.): GD 2005, LNCS 3843, pp. 505–507, 2005.

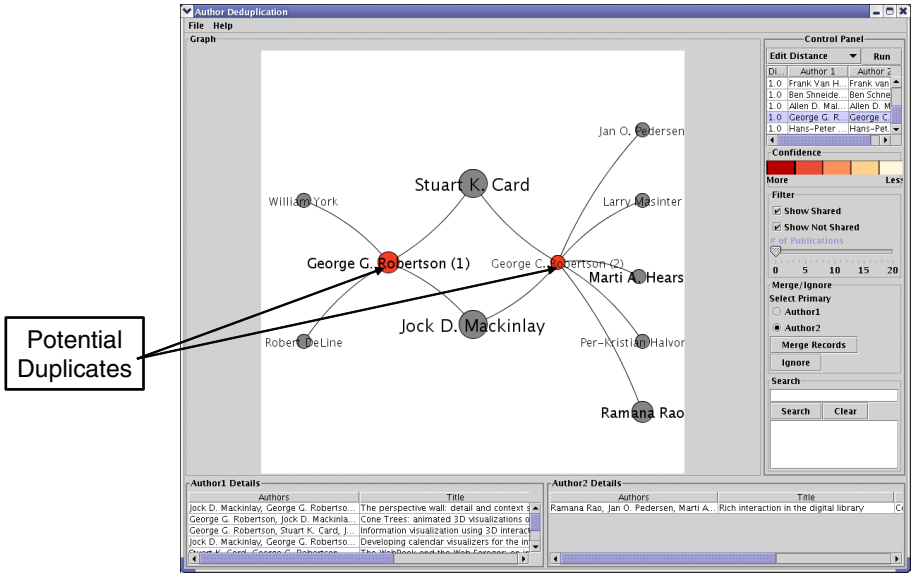

Fig. 1. The D-Dupe Interface

2 Overview

D-Dupe consists of three coordinated windows (Fig. 1). The left panel is the context collaboration graph, main controls are on the right panel, and the bottom panel displays the details on demand for the nodes.

Context Collaboration Graph (CCG): One novelty of D-Dupe is that the network visualization is tuned to the deduplication task. The CCG is the relevant subgraph of the whole network, where one duplicate pair and their immediate neighbors are displayed at a time. D-Dupe simplifies the graph further by showing only the edges between the possible duplicates and their neighbors. The potential duplicate pairs are colored in a shade of red in the tool; dark red pairs are more likely to be duplicates.

Control Panel: D-Dupe allows integration of a variety of machine learning algorithms for finding possible duplicates. After users select one of the algorithms, a table of potential duplicates is populated for users to inspect. Clicking on a potential duplicate pair will show the corresponding collaboration graph in the CCG. After inspecting the CCG, users can decide to either merge these duplicate records or disambiguate them by marking them as distinct entities. D-Dupe provides further graph filtering options such as filtering the authors based on number of publications.

Details on Demand: This panel displays descriptive information about the potential duplicate pairs; for instance in bibliographic domain, this panel dis-

plays the publications (the title, the date, the source, other authors, etc.) of the potential duplicate authors under inspection.

Acknowledgments. The work of Mustafa Bilgic, Louis Licamele, and Lise Getoor has been supported by the National Science Foundation, the National Geospatial Agency, and the UMD Joint Institute for Knowledge Discovery.