Robust Group Linkage

Pei Li 1, Xin Luna Dong 2, Songtao Guo 3, Andrea Maurino 4, Divesh Srivastava 5

1 University of Zurich
peili@ifi.uzh.ch

2 Google
lunadong@google.com

3 LinkedIn
songtao.gg@gmail.com

4 University of Milan-Bicocca
maurino@disco.unimib.it

5 AT&T Labs-Research
divesh@research.att.com

Abstract—We study the problem of group linkage: linking records that refer to entities in the same group. Applications for group linkage include finding businesses in the same chain, finding social network users from the same organization, finding players from the same team, etc. Group linkage faces new challenges compared to traditional record linkage. First, although different members in the same group can share some similar global values of an attribute, they represent different entities so can also have distinct local values for the same or different attributes, requiring a high tolerance for value diversity, while being able to distinguish local values from erroneous values in the data.

We present a robust two-stage algorithm: the first stage identifies pivots, i.e., sets of records that are very likely to belong to the same group, while being robust to possible erroneous values; the second stage collects strong evidence from the pivots and leverages it for merging more records into the same group, while being tolerant to differences in local values of an attribute. Experimental results show the high effectiveness and efficiency of our algorithm on various real-world data sets.

I. INTRODUCTION

Record linkage aims at linking records that refer to the same real-world entity and it has been extensively studied in the literature (surveyed in [1], [2]). In this paper we study a related but different problem that we call group linkage: linking records that refer to entities in the same group.

One motivation for our work comes from the need to group the millions of social network users (e.g., LinkedIn1) by their organizations, which improves matching and recommendation activities in social networks, counting authors from the same institution, and finding players of the same team. The organization information is often missing, incomplete, or simply too heterogeneous to be recognized as the same (e.g., “International Business Machines Corporation”, “IBM Corp.”, “IBM”, “IBM Research Labs”, “IBM-Almaden”, etc., all refer to the same organization). Contact phones, email addresses, and mailing addresses of people all provide extra evidence for group linkage, but they can also vary for different people even in the same organization.

We are also motivated by applications where we need to identify business chains, connected business entities that share a brand name and provide similar products and services (e.g., Walmart, McDonald’s). With the advent of the Web and mobile devices, we are observing a boom in local search; that is, searching local businesses under geographical constraints. Local search engines include Google Maps, Yahoo! Local, YellowPages, yelp, ezlocal, etc. The knowledge of business chains can have a big economic value to local search engines. However, business listings are rarely associated with specific chains explicitly stated in real-world business-listing collections. Sharing the same name, phone number, or URL domain name can all serve as evidence of belonging to the same chain. But the same value is often presented in different ways and there are many erroneous values; e.g., the domain name “farmbureauinsurance-mi.com” is incorrectly associated with 15 insurance companies, each of which has up to 8 correct domain names. We cannot easily develop any rule set that applies to all chains.

Group linkage faces new challenges compared to traditional record linkage. First, although different members in the same group can share some similar global values of an attribute, they represent different entities so can also have distinct local values for the same or different attributes. For example, different branches in the same business chain can provide different local phone numbers, different addresses, etc. Second, it is non-trivial to distinguish such differences from various representations for the same value and sometimes erroneous values in the data. We use the following example throughout the paper for illustration.

Example 1.1: We consider a set of 18M real-world business listings in the US extracted from a local search engine, each describing a business by its name, phone number, URL domain name, location, and category. Our algorithm automatically finds 600K business chains and 2.7M listings that belong to these chains. Table I lists the five largest chains we found. We observe that (1) each chain contains up to 22K different branch

<table>
<thead>
<tr>
<th>Name</th>
<th>#Store</th>
<th>#Name</th>
<th>#Phn</th>
<th>#URL</th>
<th>#Cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Starbucks</td>
<td>21,912</td>
<td>772</td>
<td>21,483</td>
<td>6</td>
<td>23</td>
</tr>
<tr>
<td>Bank of America</td>
<td>21,727</td>
<td>48</td>
<td>6,573</td>
<td>186</td>
<td>24</td>
</tr>
<tr>
<td>U-Haul</td>
<td>21,638</td>
<td>2,340</td>
<td>18,384</td>
<td>14</td>
<td>20</td>
</tr>
<tr>
<td>USPS - United State Post Office</td>
<td>19,225</td>
<td>12,345</td>
<td>5,761</td>
<td>282</td>
<td>22</td>
</tr>
<tr>
<td>McDonald’s</td>
<td>17,289</td>
<td>2,401</td>
<td>16,607</td>
<td>568</td>
<td>47</td>
</tr>
</tbody>
</table>

TABLE II
REAL-WORLD BUSINESS LISTINGS. WE SHOW ONLY STATE FOR LOCATION AND SIMPLIFY NAMES OF CATEGORY. THERE IS A WRONG VALUE IN ITALIC FONT.

<table>
<thead>
<tr>
<th>HLU</th>
<th>name</th>
<th>phone</th>
<th>URL (domain)</th>
<th>location</th>
<th>category</th>
</tr>
</thead>
<tbody>
<tr>
<td>r1</td>
<td>Home Depot, The</td>
<td>808</td>
<td></td>
<td>NJ</td>
<td>furniture</td>
</tr>
<tr>
<td>r2</td>
<td>Home Depot, The</td>
<td>808</td>
<td></td>
<td>NY</td>
<td>furniture</td>
</tr>
<tr>
<td>r3</td>
<td>Home Depot, The</td>
<td>808</td>
<td>homedepot</td>
<td>MD</td>
<td>furniture</td>
</tr>
<tr>
<td>r4</td>
<td>Home Depot, The</td>
<td>808</td>
<td>homedepot</td>
<td>AK</td>
<td>furniture</td>
</tr>
<tr>
<td>r5</td>
<td>Home Depot, The</td>
<td>808</td>
<td>homedepot</td>
<td>MI</td>
<td>furniture</td>
</tr>
<tr>
<td>r6</td>
<td>Home Depot, The</td>
<td>101</td>
<td>homedepot</td>
<td>NY</td>
<td>furniture</td>
</tr>
<tr>
<td>r7</td>
<td>Home Depot, USA</td>
<td>102</td>
<td>homedepot</td>
<td>NY</td>
<td>furniture</td>
</tr>
<tr>
<td>r8</td>
<td>Home Depot USA</td>
<td>103</td>
<td>homedepot</td>
<td>WV</td>
<td>furniture</td>
</tr>
<tr>
<td>r9</td>
<td>Home Depot USA</td>
<td>808</td>
<td></td>
<td>SD</td>
<td>furniture</td>
</tr>
<tr>
<td>r10</td>
<td>Home Depot - Tools</td>
<td>808</td>
<td></td>
<td>FL</td>
<td>furniture</td>
</tr>
<tr>
<td>r11</td>
<td>Taco Casa</td>
<td>900</td>
<td>tacocasa</td>
<td>AL</td>
<td>restaurant</td>
</tr>
<tr>
<td>r12</td>
<td>Taco Casa</td>
<td>900</td>
<td>tacocasa</td>
<td>AL</td>
<td>restaurant</td>
</tr>
<tr>
<td>r13</td>
<td>Taco Casa</td>
<td>900</td>
<td>tacocasa,</td>
<td>AL</td>
<td>restaurant</td>
</tr>
<tr>
<td>r14</td>
<td>Taco Casa</td>
<td>900</td>
<td>tacocasatexas</td>
<td>AL</td>
<td>food</td>
</tr>
<tr>
<td>r15</td>
<td>Taco Casa</td>
<td>900</td>
<td>tacocasatexas</td>
<td>AL</td>
<td>food</td>
</tr>
<tr>
<td>r16</td>
<td>Taco Casa</td>
<td>701</td>
<td>tacocasatexas</td>
<td>TX</td>
<td>restaurant</td>
</tr>
<tr>
<td>r17</td>
<td>Taco Casa</td>
<td>702</td>
<td>tacocasatexas</td>
<td>TX</td>
<td>restaurant</td>
</tr>
<tr>
<td>r18</td>
<td>Taco Casa</td>
<td>703</td>
<td>tacocasatexas</td>
<td>TX</td>
<td>restaurant</td>
</tr>
<tr>
<td>r19</td>
<td>Taco Casa</td>
<td>704</td>
<td></td>
<td>NY</td>
<td>food store</td>
</tr>
<tr>
<td>r20</td>
<td>Taco Casa</td>
<td>704</td>
<td></td>
<td>AK</td>
<td>restaurant</td>
</tr>
</tbody>
</table>

stores, (2) different branches from the same chain can have a large variety of names, phone numbers, and URL domain names, and (3) even chains of similar sizes can have very different numbers of distinct URLs (same for other attributes). Thus, rule-based linkage can hardly succeed and scalability is essential.

Table II shows 20 business listings (with some abstraction) in this data set. After investigating their webpages manually, we find that $r_1 - r_{18}$ belong to three business chains: $Ch_1 = \{r_1 - r_{10}\}$, $Ch_2 = \{r_{11} - r_{15}\}$, and $Ch_3 = \{r_{16} - r_{18}\}$; $r_{19}$ and $r_{20}$ do not belong to any chain. Note the slightly different names for businesses in chain $Ch_1$; also note that $r_{13}$ is integrated from different sources and contains two URLs, one (tacocasatexas) being wrong.

Simple linkage rules do not work well on this data set. For example, if we require only high similarity on name for chain identification, we may wrongly decide that $r_{11} - r_{20}$ all belong to the same chain as they share a popular restaurant name Taco Casa. Traditional linkage strategies do not work well either. If we apply Swosh-style linkage [3] and iteratively merge records with high similarity on name and shared phone or URL, we can wrongly merge $Ch_2$ and $Ch_3$ because of the wrong URL from $r_{13}$. If we require high similarity between listings on name, phone, URL, category, we may either split $r_6 - r_8$ out of chain $Ch_1$ because of their different local phone numbers, or learn a low weight for phone and split $r_9 - r_{10}$ out of chain $Ch_1$ since sharing the same phone number, the main evidence, is downweighted. 

The key idea in our solution is to find strong evidence that can glue group members together, while being tolerant to differences in values specific for individual group members. For example, we wish to reward sharing of primary values, such as primary phone numbers or URL domain names for chain identification, but would not penalize differences in local values, such as locations, local phone numbers, and even categories. For this purpose, our algorithm proceeds in two stages. First, we identify pivots containing records that are very likely to belong to the same group. Second, we collect strong evidence from the resulting pivots, such as primary phone numbers and URL domain names in business chains, based on which we cluster the pivots and remaining records into groups. The use of pivots and strong evidence distinguishes our clustering algorithm from traditional clustering techniques for record linkage. In this process, it is crucial that pivot generation makes very few false positives even in the presence of erroneous values, to avoid the ripple effect on clustering later. Our algorithm is designed to ensure scalability.

The group linkage problem we study in this paper is different from the group linkage in [4], [5], which compute group-level similarity between pre-specified groups of records from the same entity. Our goal is to find records of different entities that belong to the same group and we make three contributions.

1) We study pivot generation in the presence of erroneous data. Our pivot is robust in the sense that even if we remove a few possibly erroneous records from a pivot, we still have strong evidence that the rest of the records in the pivot must belong to the same group.

2) We then reduce the group linkage problem to clustering pivots and the remaining records. Our clustering algorithm leverages strong evidence collected from pivots and meanwhile is tolerant to value variety of records in the same group.

3) We conduct experiments on two real-world data sets in different domains, showing high efficiency and effectiveness of our algorithms.

Note that this paper focuses on finding records that belong to the same group. It does not require applying record linkage to identify records that refer to the same individual entity. However, we show empirically that combining our algorithm and record linkage can improve the results of both (Section VI-B3).

In the rest of the paper, Section II discusses related work. Section III defines the problem and provides an overview of our solution. Sections IV-V describe the two stages in our solution. Section VI describes experimental results. Section VII concludes.

II. RELATED WORK

Record linkage has been extensively studied in the past (surveyed in [1], [2]). Traditional linkage techniques aim at linking records that refer to the same real-world entity, so implicitly assume value consistency between records that should be linked. Group linkage is different in that it aims at linking records that refer to different entities in the same group. The variety of individual entities requires better use of strong evidence and tolerance on different values even within the same group. These two features differentiate our work from any previous linkage technique.

For record clustering in linkage, existing work may apply the transitive rule [6], do match-and-merge [3], or reduce it to an optimization problem [7]. Our work is different in that our pivot-identification algorithm aims at being robust to a few
erroneous records; and our clustering algorithm emphasizes leveraging the strong evidence collected from the pivots.

For record-similarity computation, existing work can be rule based [6], classification based [8], or distance based [9]. There has also been work on weight (or model) learning from labeled data [8], [10]. Our work is different in that in addition to learning a weight for each attribute, we also learn a weight for each value based on whether it serves as important evidence for the group. Note that some previous works are also tolerant to different values but leverage evidence that may not be available in our contexts: [11] is tolerant to schema heterogeneity from different relations by specifying matching rules; [12] is tolerant to possibly false values by considering agreement between different data providers; [13] is tolerant to out-of-date values by considering time stamps; we are tolerant to value diversity within the same group.

Two-stage clustering has been proposed in DB and IR communities [14], [15], [16], [17], [18], [19]; however, they identify pivots in different ways. Techniques in [15], [17] consider a pivot as a single record, either randomly selected or selected according to the weighted degrees of nodes in the graph. Techniques in [18] generate pivots using agglomerative clustering but can be too conservative and miss strong evidence. Techniques in [14] identify pivots as bi-connected components, where removing any node would not disconnect the graph. Although this corresponds to the 1-robustness requirement in our solution (defined in Section IV), they generate overlapping clusters; it is not obvious how to derive non-overlapping clusters in applications such as business-chain identification and how to extend their techniques to guarantee k-robustness. Techniques in [15], [16] require knowledge of the number of clusters for one of the stages, so do not directly apply in our context. Techniques in [19] generate pivots to collect evolution evidence for resolving entities that evolve over time, but may not be robust against erroneous values. We compare with these methods whenever applicable in experiments (Section VI), showing that our algorithm is robust in presence of erroneous values and consistently generates high-accuracy results on data sets with different features.

Finally, we distinguish our work from the group linkage in [4], [5], which has a different goal of identifying groups of records associated with the same entity from multiple databases. On et al. [5] compute group-level similarity between pre-specified groups of records based on record-level similarity. With the same goal, Huang [4] decided group-level similarity using network evolution analysis. Our goal is to find records of different entities that belong to the same group.

III. OVERVIEW

A. Problem Definition

Let \( R \) be a set of records that describe real-world entities by a set of attributes \( A \). For each record \( r \in R \), we denote by \( r.A \) its value of attribute \( A \in A \). Records may contain erroneous or missing values.

We consider the group linkage problem: that is, finding records that represent entities belonging to the same real-world group. As an example application, we wish to find business chains—a set of business entities with the same or highly similar names that provide similar products and services (e.g., Walmart, Home Depot, Subway and McDonald’s).\(^2\) We focus on non-overlapping groups, which often hold in applications.

Definition 3.1 (Group linkage): Given a set \( R \) of records, group linkage identifies a set of clusters \( CH \) of records in \( R \), such that (1) records that represent real-world entities in the same group belong to one cluster, and (2) records from different groups belong to different clusters.

Example 3.2: Consider the records in Example 1.1, where each record describes a business listing by attributes name, phone, URL, location, and category.

The ideal solution to the group linkage problem contains 5 clusters: \( CH_1 = \{ r_1 - r_{10} \}, CH_2 = \{ r_{11} - r_{15} \}, CH_3 = \{ r_{16} - r_{18} \}, CH_4 = \{ r_{19} \}, \) and \( CH_5 = \{ r_{20} \} \). Among them, \( CH_2 \) and \( CH_3 \) represent two different chains with the same name. \( \Box \)

B. Overview of Our Solution

Group linkage is related to but different from traditional record linkage because it essentially looks for records that represent entities in the same group, rather than records that represent exactly the same entity. Different members in the same group often share a certain amount of commonality (e.g., common name, primary phone, and URL domain of chain stores), but meanwhile can also have a lot of differences (e.g., different addresses, local phone numbers, and local URL domains); thus, we need to allow much higher variety in some attribute values to avoid false negatives. On the other hand, as we have shown in Example 1.1, simply lowering our requirement on similarity of records or similarity of a few attributes in clustering can lead to a lot of false positives.

The key intuition of our solution is to distinguish between strong evidence and weak evidence. For example, different branches in the same business chain often share the same URL domain name and those in North America often share the same 1-800 phone number. Thus, a URL domain or phone number shared among many business listings with highly similar names can serve as strong evidence for chain identification. In contrast, a phone number shared by only a couple of business entities is much weaker evidence, since one might be an erroneous or out-of-date value.

To facilitate leveraging strong evidence, our solution consists of two stages. The first stage collects records highly likely to belong to the same group; for example, a set of business listings with the same name and phone number are very likely to be in the same chain. We call the results pivots; from them we can collect strong evidence such as name, primary phone number, and primary URL domain of chains. The key goal of this stage is to be robust against erroneous values and make as few false positives as possible, so we can avoid identifying strong evidence wrongly and causing incorrect ripple effect later; however, we need to keep in mind that being too strict can miss important strong evidence.

The second stage clusters pivots and remaining records into groups according to the discovered strong evidence. It decides whether several pivots belong to the same group, and whether a record that does not belong to any pivot actually belongs to some group. It also employs weak evidence, but treats it differently from strong evidence. The key intuition of this stage is to leverage the strong evidence and meanwhile be tolerant to diversity of values in the same group so we can reduce false negatives made in the first stage.

We illustrate our approach for business-chain identification.

**Example 3.3:** Continue with the motivating example. In the first stage we generate three pivots: \( C_{r1} = \{r_1-r_7\}, C_{r2} = \{r_{14}, r_{15}\}, C_{r3} = \{r_{16}-r_{18}\} \). Records \( r_1-r_7 \) are in the same pivot because they have the same name, five of them \( (r_1-r_5) \) share the same phone number 808 and five of them \( (r_3-r_7) \) share the same URL homedepot. Similar for the other two pivots. Note that \( r_{13} \) does not belong to any pivot, because one of its URLs is the same as that of \( r_{11}-r_{12} \), and one is the same as that of \( r_{16}-r_{18} \), but except name, there is no other common information between these two groups of records. To avoid mistakes, we defer the decision on \( r_{13} \). Indeed, recall that \( \text{tacocasatexas} \) is a wrong value for \( r_{13} \). For a similar reason, we defer the decision on \( r_{12} \).

In the second stage, we generate groups – business chains. We merge \( r_8-r_{10} \) with pivot \( C_{r1} \), because they have similar names and share either the primary phone number or the primary URL. We also merge \( r_{11}-r_{13} \) with pivot \( C_{r2} \), because (1) \( r_{12}-r_{13} \) share the primary phone 900 with \( C_{r2} \), and (2) \( r_{11} \) shares the primary URL \( \text{tacocasa} \) with \( r_{12}-r_{13} \). We do not merge \( C_{r2} \) and \( C_{r3} \) though, because they share neither the primary phone nor the primary URL. We do not merge \( r_{19} \) or \( r_{20} \) to any pivot, because there is again not much strong evidence. We thus obtain the ideal result.

We describe pivot identification in Section IV and group linkage in Section V. While we present the algorithms for the setting of one machine, many components of our algorithms can be easily parallelized in Hadoop infrastructure [20]; we omit the details as it is not the focus of this paper. Before we proceed, we first introduce an important concept used in our algorithms.

**C. Attribute categorization**

We classify attributes into three categories.

- **Common-value attribute:** We call an attribute \( A \) a common-value attribute if all entities in the same group have the same or highly similar \( A \)-values (e.g., **business-name** for chain identification and **organization** for organization linkage).

- **Primary-value attribute:** We call an attribute \( A \) a primary-value attribute if entities in the same group often share one or a few primary \( A \)-values (but there can also be other values), and these values are seldom used by entities outside the group (e.g., **phone** and **URL-domain** for chain identification, and **office-address**, **phone-prefix**, and **email-server** for organization linkage).

- **Multi-value attribute:** We call the rest of the attributes multi-value attributes as there is often a many-to-many relationship between groups and values of these attributes (e.g., **category** for chain identification).

According to the definition, for each attribute \( A \) we define three measures to capture its features. First, for each group \( Ch \) and its primary \( A \)-value \( v_{Ch}^A \) (i.e., the \( A \)-value that occurs most often in \( Ch \)), we define \( m_{Ch}^A \) as the percentage of records in \( Ch \) that contain \( v_{Ch}^A \), define \( s_{Ch}^A \) as the average similarity between \( v_{Ch}^A \) and other \( A \)-values in \( Ch \), and define \( n_{Ch}^A \) as the percentage of other groups that contain \( v_{Ch}^A \). Then, we compute \( m^A, s^A, n^A \) as the average of the top-k% values; to avoid being biased by small groups, we may filter groups of small sizes\(^3\). According to the categorization, \( A \) is a common-value attribute if \( s^A \) is high; \( A \) is a primary-value attribute if \( m^A \) is high and \( n^A \) is low; and \( A \) is a multi-value attribute if \( n^A \) is high. We decide whether a measure is high or low using the largest gap between continuous measures.

Finally, we point out that despite the importance of attribute categorization, simply applying traditional record linkage methods while treating the attributes of different categories differently does not work well (Figure 3).

**IV. Pivot Identification**

The first stage of our solution creates pivots consisting of records that are very likely to belong to the same group. To this end, we only consider common-value and primary-value attributes to identify pivots. The key goal in pivot identification is to be robust to possible erroneous values. This section starts with the definition of pivots (Section IV-A), then describes how we efficiently construct similarity graphs to facilitate pivot finding (Section IV-B), and finally gives the algorithm for pivot identification (Section IV-C).

**A. Criteria for a Pivot**

At the first stage, we wish to make only decisions that are highly likely to be correct. First, we require that each pivot contains only highly similar records, and different pivots are fairly different and easily distinguishable from each other. Second, we wish that our results are robust even in the presence of a few erroneous values in the data. In the motivating example, \( r_1-r_7 \) form a good pivot, because 808 and homedepot are very popular values among these records. In contrast, \( r_{13} \) do not form a good pivot, because records \( r_{14}-r_{15} \) and \( r_{16}-r_{18} \) do not share any phone number or URL domain; the only “connector” between them is \( r_{13} \), so they can be wrongly merged if \( r_{13} \) contains erroneous values. Also, considering \( r_{13} \) as two different pivots is risky, because (1) it is not very clear whether \( r_{13} \) is in the same chain as \( r_{14}-r_{15} \) or as \( r_{16}-r_{18} \), and (2) these two pivots share one URL domain name so are not fully distinguishable.

We capture this intuition with **connectivity of a similarity graph**. We define the **similarity graph** of a set \( R \) of records as

\[ \text{similarity graph} \; \text{of} \; R \]
an undirected graph, where each node represents a record in $\mathbf{R}$, and an edge connects two nodes if they may contain strong evidence indicating a group. We consider two records sharing strong evidence if they agree on common-value attributes and (at least one) primary-value attribute. Note that our techniques are independent of the similarity criteria we apply. Figure 1 shows the similarity graph for the motivating example.

Each pivot would correspond to a connected subgraph of the similarity graph. We would like such a subgraph to be robust such that even if we remove a few nodes the subgraph remains connected; intuitively, even if there are some erroneous records, without them we still have enough evidence showing that the rest of them should belong to the same group.

**Definition 4.1 (k-robustness):** A graph $G$ is $k$-robust if after removing arbitrary $k$ nodes and edges to these nodes, $G$ is still connected. A clique or a single node is defined to be $k$-robust for any $k$.

In Figure 1, the subgraph with nodes $r_1 - r_7$ is 2-robust. It is not 3-robust as removing $r_3 - r_5$ can disconnect it.

According to the definition, we can partition the similarity graph into a set of $k$-robust subgraphs. As we do not wish to split any pivot unnecessarily, we require the maximal $k$-robust partitioning.

**Definition 4.2 (Maximal $k$-robust partitioning):** Let $G$ be a similarity graph. A partitioning of $G$ is a maximal $k$-robust partitioning if it satisfies the following properties.

1) Each node belongs to one and only one partition.
2) Each partition is $k$-robust.
3) The result of merging any partitions is not $k$-robust.

Note that a data set can have more than one maximal $k$-robust partitioning. Consider $r_{11} - r_{18}$ in Figure 1. There are three maximal 1-robust partitionings: $\{\{r_{11}\}, \{r_{12}, r_{14} - r_{15}\}, \{r_{13}, r_{16} - r_{18}\}\}$; $\{\{r_{11} - r_{12}\}, \{r_{14} - r_{15}\}, \{r_{13}, r_{16} - r_{18}\}\}$; and $\{\{r_{11} - r_{15}\}, \{r_{16} - r_{18}\}\}$. If we treat each partitioning as a possible world, records that belong to the same partition in all possible worlds have high probability to belong to the same group and so form a pivot. So, we define a pivot as follows.

**Definition 4.3 (k-Pivot):** Let $\mathbf{R}$ be a set of records and $G$ be the similarity graph of $\mathbf{R}$. The records that belong to the same subgraph in every maximal $k$-robust partitioning of $G$ form a $k$-pivot of $\mathbf{R}$. A pivot contains at least 2 records.

**Property 4.4:** A $k$-pivot is $k$-robust.

**Example 4.5:** Consider Figure 1 and assume $k = 1$. There are two connected sub-graphs. For records $r_1 - r_7$, the subgraph is 1-robust, so they form a 1-pivot. For records $r_{11} - r_{18}$, there are three maximal 1-robust partitionings for the subgraph, as we have shown. Two subsets of records belong to the same subgraph in each partitioning: $\{r_{14} - r_{15}\}$ and $\{r_{16} - r_{18}\}$; they form two 1-pivots.

The higher $k$, the stronger is the requirement for robustness; when $k = 0$, each connected subgraph is a pivot and the results would be vulnerable to erroneous values. In our motivating example, when $k = 0$, records $r_{11} - r_{18}$ would be wrongly considered as a pivot and thus belonging to the same chain. Our experiments show that a $k$ in the range $[2, 5]$ is the best, improving the results (F-measure) by 10% over $k = 0$.

### B. Constructing Similarity Graphs

Recall that we compare records by their common-value and primary-value attribute values. All records agreeing on the common-value attributes and at least one value on a primary-value attribute form a clique, which we call a v-clique. We can thus represent the similarity graph with a set of v-cliques, denoted by $C$; for example, the graph in Figure 1 can be represented by five v-cliques ($C_1 - C_5$). In addition, we maintain an inverted index $L$, where each entry corresponds to a record $r$ and contains the v-cliques that $r$ belongs to.

Given the sheer number of records in $\mathbf{R}$, the inverted index can still be huge. In fact, according to Theorem 4.6, records in the same v-clique but not any other v-clique must belong to the same pivot, so we do not need to distinguish them. Thus, we simplify the inverted index such that for each v-clique we keep only a representative for nodes belonging only to this v-clique. Table III shows the simplified index for the similarity graph in Figure 1.

**Theorem 4.6:** Let $G$ be a similarity graph and $G'$ be a graph derived from $G$ by merging nodes that belong to only one and the same v-clique. Two nodes belong to the same pivot of $G'$ if and only if they belong to the same pivot of $G$.

We can easily find v-cliques and construct the inverted index by scanning values of primary-value attributes. The running time is linear in the number of values from primary-value attributes.

**Case study:** On a data set with 18M records (described in Section VI), GRAPHCONSTRUCTION finished in 1.9 hours. The original similarity graph contains 18M nodes and 4.2B edges. The inverted index is of size 89MB, containing 3.8M entries, each associated with at most 8 v-cliques; in total there are 1.2M v-cliques. The simplified inverted index is of size...
34MB, containing 1.5M entries, where an entry can represent up to 11K records. The simplified inverted index reduces the size of the similarity graph by 3 orders of magnitude.

C. Identifying Pivots

We solve the pivot-identification problem by reducing it to a Max-flow/Min-cut Problem. However, computing the max flow for a given graph \( G \) and a source-destination pair takes time \( O(|G|^{2.5}) \), where \(|G|\) denotes the number of nodes in \( G \); even the simplified inverted index can still contain millions of entries, so it can be very expensive. We thus first screen graph \( G \) in two steps (SCREEN in Section IV-C1): (1) merging certain v-cliques according to a sufficient (but not necessary) condition for \( k \)-robustness and consider them as a whole in pivot identification; then (2) splitting \( G \) into subgraphs according to a necessary (but not sufficient) condition for \( k \)-robustness. We find the max flow only on the resulting subgraphs, which are substantially smaller (SPLIT in Section IV-C2). Section IV-C3 gives the full algorithm, which iteratively applies SCREEN and SPLIT.

1) Screening: We explain how to reduce the search space for Max-flow by screening. Experiments show that we reduce the input size by 4 orders of magnitude on a real-world data set of 18M records.

A graph can be considered as a union of v-cliques, so essentially we need to decide if a union of v-cliques is \( k \)-robust. First, we can prove the following sufficient condition for \( k \)-robustness.

**Theorem 4.7 ((\( k + 1 \))-connected condition):** Let \( G \) be a graph consisting of a union \( Q \) of v-cliques. If for every pair of v-cliques \( C, C' \in Q \), there is a path of v-cliques between \( C \) and \( C' \) and every pair of adjacent v-cliques on the path share at least \( k + 1 \) nodes, graph \( G \) is \( k \)-robust.

Given Theorem 4.7, we define a \((k + 1)\)-connected v-union.

**Definition 4.8 ((\( k + 1 \))-connected v-union):** A \((k + 1)\)-connected v-union is a union of v-cliques that satisfies \((k + 1)\)-connected condition.

A \((k + 1)\)-connected v-union (simplified as v-union) must be \( k \)-robust but not vice versa. In Figure 1, subgraph \( \{r_1 - r_7\} \) is a 3-connected v-union, because the only two v-cliques, \( C_1 \) and \( C_2 \), share 3 nodes. Indeed, it is 2-robust. On the other hand, graph \( G_1 \) in Figure 2 is 2-robust but not 3-connected (there are 4 v-cliques, where each pair of adjacent v-cliques share only 1 or 2 nodes). From Theorem 4.7, we have that a \((k + 1)\)-connected v-union is \( k \)-robust and accordingly a \( k \)-pivot. Therefore for pivot identification we consider graph \( G \) as a set of v-unions instead of a set of v-cliques. Figure 8 shows that this improves the runtime for pivot identification by 47% on a real-world data set of 11M records.

Next, we present a necessary condition for \( k \)-robustness.

**Theorem 4.9 ((\( k + 1 \))-overlap condition):** Graph \( G \) is \( k \)-robust only if for every \((k + 1)\)-connected v-union \( Q \in G \), \( Q \) shares at least \( k + 1 \) common nodes with the subgraph consisting of the rest of the v-unions.

Accordingly, we define a \((k + 1)\)-overlap graph.

**Definition 4.10 ((\( k + 1 \))-overlap Graph):** A \((k + 1)\)-overlap graph is a graph that satisfies \((k + 1)\)-overlap condition.

A \( k \)-robust graph must be a \((k + 1)\)-overlap graph but not vice versa. In Figure 1, subgraph \( \{r_{11} - r_{18}\} \) is a 2-overlap graph, because there are two 2-connected v-unions, \( \{r_{11} - r_{15}\} \) and \( \{r_{13}, r_{16} - r_{18}\} \), but they share only one node; indeed, the subgraph is not \( 1 \)-robust. On the other hand, graph \( G_2 \) in Figure 2 satisfies the 3-overlap condition, as it contains four 3-connected v-unions (actually four v-cliques), \( Q_1 - Q_4 \), and each v-union shares 3 nodes in total with the others; however, it is not \( 2 \)-robust (removing \( r_3 \) and \( r_4 \) disconnects it). According to Theorem 4.9, if graph \( G \) is not a \((k + 1)\)-overlap graph, it cannot be \( k \)-robust. Therefore we split \( G \) into a set of \((k + 1)\)-overlap subgraphs, and check \( k \)-robustness by Max-flow on each subgraph. Figure 8 shows that it further improves the runtime by at least three orders of magnitude.

Now the problem is to split \( G \) into \((k + 1)\)-overlap subgraphs. Let \( G \) be a graph where a \((k + 1)\)-connected v-union overlaps with the rest of the v-unions on no more than \( k \) nodes. We split \( G \) by removing these overlapping nodes. For subgraph \( \{r_{11} - r_{18}\} \) in Figure 1, we remove \( r_{13} \) and obtain two subgraphs \( \{r_{11} - r_{12}, r_{14} - r_{15}\} \) and \( \{r_{16} - r_{18}\} \) (recall from Example 4.5 that \( r_{13} \) cannot belong to any pivot). Note that the result subgraphs may not be \((k + 1)\)-overlap graphs (e.g., \( \{r_{11} - r_{12}, r_{14} - r_{15}\} \)) contains two v-unions that share only one node), so we need to further screen them.

We now describe our screening algorithm, SCREEN, which takes a graph \( G \), represented by a set \( C \) of v-cliques and inverted list \( \bar{L} \), as input, finds \((k + 1)\)-connected v-unions in \( G \) and meanwhile decides if \( G \) is a \((k + 1)\)-overlap graph. If not, it splits \( G \) into subgraphs for further examination.

1) If \( G \) contains a single node, output it as a pivot if it represents records belonging only to one v-clique.
2) For each v-clique \( C \in C \), initialize a v-union \( Q(C) \). We denote the set of v-unions by \( \bar{Q} \), and the common nodes of \( C \) and \( C' \) by \( B(C, C') \).
3) For each v-clique \( C \in C \), we merge v-unions as follows.
   a) For each not-processed record \( r \in C \), for every pair of v-cliques \( C_1 \) and \( C_2 \) in \( r \)’s index entry, if they belong to different v-unions, add \( r \) to \( B(C_1, C_2) \).
   b) For each v-union \( Q \in \bar{Q} \) where there exist \( C_1 \in Q \) and \( C_2 \in B(C, C') \) such that \( |B(C_1, C_2)| \geq k + 1 \), merge \( Q \) and \( Q(C) \).
   At the end, \( \bar{Q} \) contains all \((k + 1)\)-connected v-unions.
4) For each v-union \( Q \in \bar{Q} \), find its border nodes as \( B(Q) = \cup_{C \in Q, C' \notin Q}B(C, C') \). If \( |B(Q)| \leq k \), split the subgraph it belongs to, denoted by \( G(Q) \), into two subgraphs \( Q \setminus B(Q) \) and \( G(Q) \setminus Q \).
5) Return the remaining subgraphs.

**Proposition 4.11:** Denote by $|L|$ the number of entries in input $L$. Let $m$ be the maximum number of blocks a record belongs to, and $a$ the maximum number of adjacent v-unions a v-union has. Algorithm SCREEN finds $(k + 1)$-overlap subgraphs in time $O((m^2 + a) |L|)$ and the result is independent of the order in which we examine the v-cliques.

Note that $m$ and $a$ are typically very small, so SCREEN is basically linear in the size of the inverted index. According to Theorem 4.12, we further simplify $G$ by keeping for each v-union a representative for all nodes that only belong to it.

**Theorem 4.12:** Given similarity graph $G$, let $G'$ be a graph derived from $G$ by merging nodes that belong to only one and the same v-union. Two nodes belong to the same pivot of $G'$ if and only if they belong to the same pivot of $G$.

**Example 4.13:** Consider Table III as input and $k = 1$. Step 2 creates five v-unions $Q_1 - Q_5$ for the input v-cliques.

Step 3 starts with v-clique $C_1$. It has 4 nodes (in the simplified inverted index), among which 3 are shared with $C_2$. Thus, $B(C_1, C_2) = \{t_3 - t_5\}$ and $|B(C_1, C_2)| \geq 2$, so we merge $Q_1$ and $Q_2$ into $Q_{1/2}$. Examining $C_2$ reveals no other shared node.

Step 3 then considers v-clique $C_3$. It has three nodes, among which $t_12 - t_13$ are shared with $C_4$ and $t_14$ is also shared with $C_5$. Thus, $B(C_3, C_4) = \{t_12 - t_13\}$ and $B(C_3, C_5) = \{t_11\}$. We merge $Q_3$ and $Q_4$ into $Q_{3/4}$. Examining $C_4$ and $C_5$ reveals no other shared node. We thus obtain three 2-v-unions; $Q = \{Q_{1/2}, Q_{3/4}, Q_5\}$.

Step 4 then considers each v-union. For $Q_{1/2}$, $B(Q_{1/2}) = \emptyset$ and we thus split subgraph $Q_{1/2}$ out and merge all of its nodes to one $r_{11} \ldots r_{17}$. For $Q_{3/4}$, $B(Q_{3/4}) = \{r_{13}\}$ so $|B(Q_{3/4})| < 2$. We split $Q_{3/4}$ out and obtain $\{r_{11} - r_{12}, r_{14}, r_{15}\}$ (at which we stop). Similar for $Q_5$ and we obtain $\{r_{16}, r_{17}\}$. Therefore, we return three subgraphs for further screening.

2) **Reduction:** Each result k-overlap subgraph is typically very small, and as we soon show, in practice the majority of them are already k-robust. Thus, in many cases SCREEN is adequate in finding pivots. For completeness, we next briefly describe SPLIT that guarantees k-robustness of the pivots.

Between two nodes $a, b$ in $G$, consider the paths that do not share any node except $a$ and $b$. We denote the maximal number of such paths between $a$ and $b$ by $\kappa(a, b)$. According to Menger’s Theorem [21], $\kappa(a, b)$ is the minimum number of nodes removing which disconnects $a$ and $b$. Obviously, $G$ is k-robust if $\kappa(a, b) > k$ for any nodes $a, b$ in $G$.

Note that for two nodes $a, b$ in a $(k + 1)$-connected v-union, we have $\kappa(a, b) \geq k + 1$. Theorem 4.14 further gives k-robustness condition of $G$ on adjacent v-unions in $G$.

**Theorem 4.14 (k-robustness condition):** Let $G$ be a similarity graph. Graph $G$ is k-robust if and only if for every pair of adjacent $(k + 1)$-connected v-unions $Q$ and $Q'$, there exist two nodes $a \in Q \ast Q'$ and $b \in Q' \ast Q$ such that $\kappa(a, b) > k$.

Computing $\kappa(a, b)$ is reduced to a Max-flow/Min-cut Problem in [22], for which we apply the well-known Ford & Fulkerson Algorithm [23]. If a graph $G$ is not k-robust, we shall get the set $S$ of nodes that separate $G$, which we call separator nodes. Suppose the separator nodes separate $G$ into $X$ and $Y$ (there can be more subgraphs). We split $G$ into $X \cup S$ and $Y \cup S$ for further processing.

Note that we need to include $\bar{S}$ in both sub-graphs to maintain the integrity of each v-union. To understand why, consider $G_2$ in Figure 2 where $\bar{S} = \{r_3, r_4\}$. According to Definition 4.3, there is no 2-pivot. If we split $G_2$ into $\{r_1 - r_4\}$ and $\{r_5 - r_8\}$ (without including $\bar{S}$), both subgraphs are 2-robust and we would return them as 2-pivots. The problem happens because v-cliques $Q_1 - Q_4$ “disappear” after we remove the separators $r_3$ and $r_4$. Thus, we should split $G_2$ into $\{r_1 - r_4\}$ and $\{r_3 - r_6\}$ instead and that would further trigger splitting on both subgraphs. Eventually we wish to exclude the separator nodes from any pivot, so we mark them as “separators” and exclude them from the returned pivots.

Algorithm SPLIT takes a $(k + 1)$-overlap subgraph $G$ as input and decides if $G$ is k-robust. If not, it splits $G$ into subgraphs on which we will then re-apply screening.

1) For each pair of adjacent $(k + 1)$-connected v-unions $Q, Q' \in G$, find $a \in Q \ast Q', b \in Q' \ast Q$. Apply Ford & Fulkerson Algorithm [23] to compute $\kappa(a, b)$.

2) Once we find nodes $a, b$ where $\kappa(a, b) \leq k$, find separator nodes $\bar{S}$ that separate $G$. Remove $\bar{S}$ and obtain several subgraphs. Add $\bar{S}$ back to each subgraph and mark $\bar{S}$ as “separators”. Return the subgraphs for screening.

3) Otherwise, $G$ is k-robust and output it as a k-pivot.

**Example 4.15:** Consider graph $G_1$ in Figure 2 and $k = 2$. There are four 3-connected v-unions (actually four v-cliques) and six pairs of adjacent v-unions. For $Q_1$ and $Q_2$, we check nodes $r_2$ and $r_4$ and find $\kappa(r_2, r_4) = 3$. Similarly we check for every other pair of adjacent v-unions and decide that the graph is 2-robust.

Now consider graph $G_2$ and $k = 2$. There are four 3-connected v-unions. When we check $r_1 \in Q_1$ and $r_6 \in Q_3$, we find $\bar{S} = \{r_3, r_4\}$. We then split $G_2$ into subgraphs $\{r_1 - r_4\}$ and $\{r_3 - r_6\}$, marking $r_3$ and $r_4$ as “separators”. Remember that if we solve the Max-Flow Problem directly for each pair of sources in the original graph, the complexity is $O(g^{4.5})$, which would be dramatically higher.

3) **Full Algorithm:** We now present the full algorithm, PIVOT IDENTIFICATION (Algorithm 1). It first initializes the working query $Q$ with only input $G$ (Line 1). Each time it pops a subgraph $G'$ from $Q$ and invokes SCREEN (Lines 3-4). If the output of SCREEN is still $G'$ (so $G'$ is a $(k + 1)$-overlap subgraph) (Line 5), it invokes SPLIT on $G'$ (Line 6). Subgraphs output by SCREEN and SPLIT are added to the queue for further examination (Lines 7, 9) and identified pivots are added to pivot set $\bar{C}$. It terminates when $Q = \emptyset$.

**Theorem 4.17:** Let $G$ be the input graph and $q$ be the number of $(k + 1)$-connected v-unions in $G$. Define $a, p, g, m,$ and $|L|$ as in Propositions 4.11 and 4.16. Algorithm PIVOT IDENTIFICATION finds correct pivots of $G$ in time $O(q((m^2 + \alpha)|L| + pg^{2.5}))$ and is order independent.
Method

S

Pivot

G

S

G

Pivot

S

3

58 nodes (8 v-unions) and on average 10 nodes (2.7 v-unions).

are fewer than 10 nodes and running S

SCREEN

an input contains at most 39.3K nodes; for 97% inputs there

Case study:

2

a single node that represents multiple records. It further splits

records, which together we call elements, into groups. We

apply an efficient state-of-the-art hill-climbing algorithm for

clustering (details in [24]). Note that in this stage we also con-

sider multi-value attributes that may contain weak evidence.

To avoid merging records based only on weak evidence, we

require that every group with more than one record must con-

tain a pivot. Comparing with clustering in traditional record

linkage, we make two changes in element-cluster similarity

computation. First, we identify strong evidence according to

popularities of attribute values within a group such that

similarity on primary values (strong evidence) is rewarded

more. Second, we allow diverse local values from different

entities in the same group.

Collecting strong evidence: We identify strong evidence as

poplar values within a cluster. When we maintain the

signature for a pivot or a cluster, we keep all values of an

attribute and assign a high weight to a popular value.

Specifically, let \( R \) be a set of records. Consider value \( v \) and let

\( R(v) \subseteq R \) denote the records in \( R \) that contain \( v \). The weight

of \( v \) is computed by \( w(v) = \frac{|R(v)|}{|R|} \).

Example 5.1: Consider phone for pivot \( Cr_1 = \{r_1 - r_7\} \) in

Table II. There are 7 business listings in \( Cr_1 \), 5 providing 808

(\( r_1 - r_5 \)), one providing 101 (\( r_6 \)), and one providing 102 (\( r_7 \)).

Thus, the weight of 808 is \( \frac{5}{7} \cdot .71 \) and the weight for 101 and

102 is \( \frac{1}{7} \cdot .14 \), showing that 808 is the primary phone.

Allowing diverse values: When we compute the similarity

between an element \( e \) and a cluster \( Ch \), we consider all

primary-value attributes together. To compute primary-value

attribute similarity, denoted by \( sim_{prm}(e, Ch) \), we reward

sharing primary values (values with a high weight) but not

penalizing different values, unless there is no shared value.

Specifically, if the primary value of \( e \) is the same as that of

\( Ch \), we consider them having probability \( p \) to be in the

same group. Since we use weights to measure whether the

value is primary and allow slight difference on values, with

a value \( v \) from \( e \) and \( v' \) from \( Ch \), the probability becomes

\( p \cdot w_e(v) \cdot w_{ch}(v') \cdot s(v, v') \), where \( w_e(v) \) measures the weight

of \( v \) in \( e \), \( w_{ch}(v') \) measures the weight of \( v' \) in \( Ch \), and

\( s(v, v') \) measures the similarity between \( v \) and \( v' \).

We compute \( sim_{prm}(e, Ch) \) as the probability that they

belong to the same group given several shared values.

\[
sim_{prm}(e, Ch) = 1 - \prod_{v, v' \in Ch} (1 - p \cdot w_e(v) \cdot w_{ch}(v') \cdot s(v, v')). \tag{1}
\]

When there is no shared primary value, \( sim_{prm} \) can be close

to 0; once there is such value, \( sim_{prm} \) can be significantly

increased, since we typically set a large \( p \).

Example 5.2: Consider element \( e = r_8 \) and cluster \( Ch_1 = \{r_1 - r_7\} \) in

Example 1.1, where phone and domain are

primary-value attributes. Assume \( p = .9 \). Element \( e \) and \( Ch_1 \)

share the primary email domain, with weight 1 and \( \frac{7}{9} \cdot .71 \) respectively, but have different phone numbers (assuming

similarity of 0). We compute \( sim_{prm}(e, Ch_1) = 1 - (1 - .9 \cdot 

1 \cdot .71 \cdot 1 \cdot (1 - 0) \cdot (1 - 0) \cdot (1 - 0) = .639 \); essentially, we do

not penalize the difference in phone numbers. Note however if

homedepot appeared only once, so was not a primary value, its

weight would be .14 and accordingly \( sim_{dom}(e, Ch_1) = .126 \),

indicating a much lower similarity.

V. GROUP LINKAGE

The second stage clusters the pivots and the remaining

records, which together we call elements, into groups. We

<table>
<thead>
<tr>
<th>Input</th>
<th>Method</th>
<th>Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G_1 )</td>
<td>SCREEN</td>
<td>( G_1 )</td>
</tr>
<tr>
<td>( G_1 )</td>
<td>SPLIT</td>
<td>Pivot ( G_1 = {r_1 - r_6} )</td>
</tr>
</tbody>
</table>

Algorithm 1: PivotIdentification \((G, k)\)

Input: \( G \): Simplified similarity graph, represented by \( C \) and \( L \).

\( k \): Robustness requirement.

Output: \( C \): Set of pivots in \( G \).

1: Let \( Q = \{G\}, C = \emptyset \);  
2: while \( Q \neq \emptyset \) do  
3: Pop \( G' \) from \( Q \);  
4: Let \( \bar{P} = \) SCREEN \((G', k, C)\);  
5: if \( \bar{P} = \{G'\} \) then  
6: Let \( S = \) SPLIT \((G', k, C)\);  
7: add graphs in \( S \) to \( Q \);  
8: else  
9: add graphs in \( \bar{P} \) to \( Q \);  
10: end if  
11: end while

TABLE IV


Example 4.18: First, consider graph \( G_1 \) in Figure 2 and

\( k = 2 \). Table IV shows the step-by-step pivot identification

process. It passes both screening and the \( k \)-robustness checking

in SPLIT, so the whole graph is output as a pivot.

Next, consider the motivating example, with the input shown

in Table III and \( k = 1 \). Originally, \( Q = \{G\} \). After invoking

SCREEN on \( G \), we obtain three subgraphs \( G^1, G^2, \) and \( G^3 \).  
SCREEN outputs \( G^1 \) and \( G^3 \) as 1-pivots since each contains

a single node that represents multiple records. It further splits

\( G^2 \) into two single-node graphs \( G^4 \) and \( G^5 \), and outputs the

latter as a 1-pivot.

Case study: On the data set with 18M records, we found 3-

pivots in 5.6 minutes. SCREEN was invoked 14K times and

took 51 seconds (15%) in total. Except the original graph,

an input contains at most 39.3K nodes; for 97% inputs there

are fewer than 10 nodes and running SCREEN was very fast.

SPLIT was invoked only 41 times; an input contains at most

58 nodes (8 v-unions) and on average 10 nodes (2.7 v-unions).

Recall that the simplified inverted index contains 1.5M entries,

so SCREEN reduced the size of the input to SPLIT by 4 orders

of magnitude. Only 1 input graph to SPLIT did not pass the

3-robustness check.
TABLE V

<table>
<thead>
<tr>
<th>#Records</th>
<th>#Groups</th>
<th>Group size</th>
<th>#Singletons</th>
<th>Level of distinction</th>
</tr>
</thead>
<tbody>
<tr>
<td>BizLow</td>
<td>2446</td>
<td>1</td>
<td>2446</td>
<td>0</td>
</tr>
<tr>
<td>BizAvg</td>
<td>2062</td>
<td>30</td>
<td>[2, 308]</td>
<td>503</td>
</tr>
<tr>
<td>BizHigh</td>
<td>1149</td>
<td>14</td>
<td>[33, 259]</td>
<td>0</td>
</tr>
<tr>
<td>SIGMOD</td>
<td>590</td>
<td>71</td>
<td>[2, 41]</td>
<td>162</td>
</tr>
</tbody>
</table>

VI. EXPERIMENTAL EVALUATION

This section describes experimental results on two real-world data sets. We show that (1) identifying pivots and leveraging strong evidence learnt from the pivot is crucial in group linkage (Section VI-B). (2) For pivot identification, requiring $k$-robustness with $k \in [1, 5]$ already performs well on most data sets that have reasonable number of errors (Section VI-B1). (3) For clustering, setting weights of values according to their popularity and allowing diverse values is critical for obtaining good clustering results (Section VI-B2). (4) Our algorithm is scalable (Section VI-C).

A. Experiment Settings

Data and gold standard: We experimented on two real-world data sets. Biz contains 18M US business listings and each listing has attributes name, phone, URL, location and category; we decide which listings belong to the same business chain. SIGMOD contains records of about 590 attendees of SIGMOD '98 and each record has attributes name, affiliation, address, phone, fax and email; we decide which attendees belong to the same institute.

We experimented on the whole Biz data set to study scalability of our techniques. We evaluated accuracy of our techniques on four subsets of data with different levels of distinctions (see in Table V). The first three are from Biz. (1) BizLow contains 2446 listings for the same business chain Allstate Insurance. These listings have the same name, but 1499 provide URL “allstate.com”, 854 provide another URL “allstateagencies.com”, while 130 provide both, and 227 listings do not provide any value for phone or URL. (2) BizAvg contains 2062 listings from Biz, where 1559 belong to 30 randomly selected business chains, and 503 do not belong to any chain; among the 503 listings, 86 are highly similar in name to listings in the business chains and the rest are randomly selected. (3) BizHigh data set contains 1149 listings with similar names and highly similar category values; they belong to 14 different chains. Among the listings, 708 provide the same wrong name Texas Farm Bureau Insurance and meanwhile provide a wrong URL farmbureauinsurance-mi.com. Among these three subsets, the latter two are hard cases; for each data set, we manually verified all the chains by checking store locations provided by the business-chain websites and used it as the gold standard. The last “subset” is actually the whole SIGMOD data set. It has very few wrong values, but the same affiliation can be represented in various ways and some affiliation names can be very similar (e.g., UCSC vs. UCSD). We manually identified 71 institutes that have multiple attendees and there are 162 attendees who do not belong to any of these institutes.

Measure: We considered each group as a cluster and compared pairwise linking decisions with the gold standard. We measured the quality of the results by precision ($P$), recall ($R$), and F-measure ($F$). If we denote the set of true-positive pairs by $TP$, the set of false-positive pairs by $FP$, and the set of false-negative pairs by $FN$, then, $P = \frac{|TP|}{|TP| + |FP|}$, $R = \frac{|TP|}{|TP| + |FN|}$, $F = \frac{2PR}{P + R}$. We also reported execution time.

Implementation: We implemented the technique we proposed in this paper, and call it Group. Before applying Group, we merge records that share highly similar values on pre-selected attributes (i.e., name, city and address for biz). In pivot generation, by default we considered two records are similar for Biz if (1) their name similarity is above .95; and (2) they share at least one phone or URL domain name. For SIGMOD we require (1) affiliation similarity is above .95; and (2) they share at least one phone prefix (3-digit), fax prefix (3-digit), or email server. We required 2-robustness for pivots. In clustering, (1) for blocking, we put records whose name similarity is above .8 in the same block; (2) for similarity computation, we computed string similarity by Jaro-Winkler distance [25], we set $p = .8$, and we learned attribute weights from 1000 records randomly selected from BizAvg data for Biz, and 300 records randomly selected from SIGMOD.

For comparison, we implemented the following baselines:

- **SameName** groups Biz records with highly similar names and groups SIGMOD records with highly similar affiliations (similarity above .95);
- Traditional machine-learning methods include PARTITION, CENTER and MERGE [7]; each computes record similarity as weighted sum of attribute similarities with learnt attribute weights, and applies a state-of-the-art clustering algorithm to find groups.
- Two-stage method YOSHIDA [18] generates pivots by agglomerative clustering with threshold .9 in the first stage, uses TF/IDF weights for features and applies linear algebra to cluster records in the second stage.

We implemented the algorithms in Java. We used a Linux machine with Intel Xeon X5550 processor (2.66GHz, cache 8MB, 6.4GT/s QPI) and 8GB main memory. We used MySQL to store the data and stored the index as a database table. Note that after blocking, we can fit each block of nodes or elements in memory, which is typically the case with a good blocking strategy.

B. Evaluating Effectiveness

We first evaluate effectiveness of our algorithms. Figure 3 compares Group with the baseline methods, where for the three traditional linkage methods we plot only the best results. On BizHigh, all methods put all records in the same chain because a large number (708) of listings have both a wrong name and a wrong URL. We manually perturbed the data as follows: (1) among the 708 listings with wrong URLs,

Element-cluster similarity is then computed as a weighted sum of attribute similarities (details in [24]).
408 provide a single (wrong) URL and we fixed it; (2) for all records we set name to “Farm Bureau Insurance”, so removed hints from business names. Even after perturbing, this data set remains the hardest and we use it hereafter instead of the original one for other experiments.

We observe that (1) GROUP obtains the highest F-measure (above .9) on each data set. It has the highest precision most of the time as it applies pivot identification and leverages the strong evidence collected from resulting pivots. It also has a very high recall (mostly above .95) on each subset because the clustering phase is tolerant to diversity of values within chains. (2) The F-measure of SAMENAME is up to 80% lower than GROUP. It can have false positives when listings of highly similar names belong to different chains and can also have false negatives when some listings in a chain have fairly different names from other listings. It only performs well in the easiest dataset BizLow, where it happens that all listings have the same name and belong to the same chain. (3) The highest F-measure of traditional linkage methods can be 16-41% lower than SAMENAME. It requires higher similarity than sharing name values. As a result, it has a lower recall than SAMENAME. (4) YOSHIDA has comparable precision to GROUP since its first stage is conservative too, which makes it often improve over the best of traditional linkage methods on Biz dataset where reducing false positives is a big challenge; on the other hand, its first stage is often too conservative (requiring high record similarity) so the recall is 10-35% lower than GROUP, which also makes it perform worse than traditional linkage methods on SIGMOD dataset where reducing false negatives is challenging.

**Contribution of different components:** We compared GROUP with (1) PIVOT, which conducts only pivot identification, and (2) CLUSTER, which considers each individual record as a pivot (in the spirit of [15], [17]) and conducts only clustering. Figure 4 shows the results. First, we observe that PIVOT improves over traditional linkage methods on precision by up to 79% but has a lower recall (up to 34% lower) most of the time, because it sets a high requirement for merging records into groups. Note however that its goal is indeed to obtain a high precision such that the strong evidence collected from the pivots are trustworthy for the clustering phase. Second, CLUSTER often has higher precision (by up to 77%) but lower recall (by up to 32%) than the best traditional linkage methods; their F-measures are comparable on each data set. On BizAvg it can obtain an even higher precision than PIVOT, because PIVOT can make mistakes when too many records have erroneous values, but CLUSTER may avoid some of these mistakes by considering also weak evidence for similarity. However, applying clustering on the results of CLUSTER would not change the results, but applying clustering on the results of PIVOT can obtain a much higher F-measure, especially a higher recall (98% higher than CLUSTER on BizAvg). This is because the result of CLUSTER lacks the strong evidence collected from high-quality pivots so the final results would be less tolerant to diversity of values, showing the importance of pivot identification. Finally, we observe that GROUP obtains the best results in most of the data sets.

We next evaluate various choices in the two stages. Unless specified otherwise, we observed similar patterns on all data sets, and report the results on BizAvg or perturbed BizHigh data, whichever has more distinguishable results.

1) **Pivot Identification:** **Pivot generation:** We first compared three pivot-generation strategies: PIVOT iteratively invokes SCREEN and SPLIT, ONLYSCREEN only iteratively invokes SCREEN, and YOSHIDAI generates pivots by agglomerative clustering [18]. Recall that by default we apply PIVOT. Figure 5 compares them on the perturbed BizHigh data. First,
we observe similar results of ONLYSCREEN and PIVOT on all data sets since most inputs to SPLIT pass the $k$-robustness test. Thus, although SCREEN in itself cannot guarantee soundness of the resulting pivots, it already does well in practice. Second, YOSHIDAII has lower recall in both pivot and clustering results, since it has stricter criteria in pivot generation.

**Robustness requirement:** We next studied how the robustness requirement can affect the results (Figure 6). We have three observations. (1) When $k = 0$, we essentially take every connected subgraph as a pivot, so the generated pivots can have a much lower precision; those false positives cause both a low precision and a low recall for the resulting chains because we do not collect high-quality strong evidence. (2) When we vary $k$ from 1 to 4, the number of false positives decreases while that of false negatives increases for the pivots, and the F-measure of the chains increases but only very slightly. (3) When we continue increasing $k$, the results of pivots and clusters remain stable. This is because setting $k=4$ already splits the graph into single v-cliques, so further increasing $k$ would not change the pivots. This shows that considering $k$-robustness is important, but $k$ does not need to be too high.

**Graph generation:** We compared three edge-adding strategies for similarity graphs: SIM takes weighted similarity on each attribute except location and requires a similarity of over .8; TWO DOM requires sharing name and at least two values on primary-value attributes; ONE DOM requires sharing name and one value on primary-value attributes. Recall that by default we applied ONE DOM. We observe that (1) SIM requires similar records so has a high precision, with a big sacrifice on recall for the pivots (.00025); as a result, the F-measure of the chains is very low (.59); (2) TWO DOM has the strongest requirements and so has even lower recall than SIM for the pivots (.00002), and in turn it has the lowest F-measure for the chains (.52). This shows that only requiring high precision for pivots with big sacrifice on recall can also lead to low F-measure for the chains.

2) **Clustering:** **Clustering strategy:** We first compared our clustering algorithm with two algorithms proposed for the second stage of two-stage clustering: LIU [16] iteratively applies majority voting to assign each record to a cluster and collects a set of representative features for each cluster using a threshold (we set it to 5, which leads to the best results); YOSHIDA II [18] is the second stage of YOSHIDA. Figure 7 compares their results. We observe that our clustering method improves the recall by 39% over LIU and by 11% over YOSHIDAII. LIU may filter strong evidence by the threshold; YOSHIDAII cannot handle records with null values well.

**Value weight:** We then compared the results with and without setting popularity weights for values. We observe that setting the popularity weight helps distinguish primary values from unpopular values, thus can improve the precision. Indeed, on perturbed BizAvg data it improves the precision from .11 to .98, and improves the F-measure by 403%.

**Attribute weight:** We next considered our attribute weight learning strategy. We first compared SEPARATEDPRIM ARY, which learns separated weights for different primary-value attributes, and UNITEDPRIM ARY (our default), which considers all such attributes as a whole and learns one single weight for them. On BizAvg the latter improves over the former by 95% on recall and obtains slightly higher precision, because it penalizes only if neither phone nor URL is shared and so is more tolerant to different values for primary-value attributes.

**Robustness w.r.t. parameter $p$:** We also ran experiments to test robustness against parameter setting. We observed very similar results when we ranged $p$ from .8 to 1.

3) **Preprocessing:** **Attribute categorization:** We studied attribute categorization on both data sets. For Biz we used a labeled data set of 2062 records and identified name as a common-value attribute ($m = .94, s = .85, n = .001$); for SIGMOD we used a labeled data set of 369 records and identified affiliation as a common-value attribute ($m = .77, s = .85, n = 0$).

Then for Biz we used a set of randomly selected 166,236 records and for SIGMOD we used all records. We considered only groups of size above 10; we averaged $m, s, n$ among top 1% values for Biz and among all values for SIGMOD. For Biz data, we identify URL domain and phone as primary-value attributes, and state and category as multi-value attributes. For SIGMOD data, we identify phone-prefix, fax-prefix and email-domain as primary-value attributes, and state as a multi-value attribute.

**Record linkage vs group linkage:** We studied the interaction between record linkage and group linkage on BizAvg. We compared three strategies: NO-RL does not conduct record linkage; PRE-RL (the default) conducts record linkage before group linkage; FULL-RL first conducts record linkage (merging records with highly similar name, address, city and the same phone or URL domain), then conducts group linkage, while removing duplicates within pivots or groups if they have highly similar address and city. We have three observations. First, the three approaches obtain very similar
results (F-measure all above .95). Second, applying record linkage improves group linkage but very slightly (F-measure improving from .9639 to .9643). Third, among the 2062 records in BizAvg, Pre-RL identifies 21 pairs of duplicates while Full-RL identifies 59 pairs, showing that group linkage can improve record linkage as well.

C. Evaluating Efficiency

Our algorithm finished in 8.3 hours on Biz data set with 18M listings on a single machine. Note that simpler methods (which we describe shortly) took over 10 hours for Stage I on 20% of the same data set. Also note that using Hadoop can reduce execution time for graph construction from 1.9 hours to 37 minutes; we skip the details as it is not the focus of the paper.

Stage I: It spent 1.9 hours for graph construction and 2.2 minutes for pivot generation. To test scalability, we randomly divided the data set into five subsets of the same size; we started with one subset and gradually added the others. We compared five pivot generation methods: Naïve applies Split on the original graph; INDEX optimizes Naïve by using an inverted index; SINDEX simplifies the inverted list by Theorem 4.6; Union in addition merges v-cliques into v-unions by Theorem 4.7; Pivot in addition splits the input graph by Theorem 4.9. Figure 8 shows the results and we have five observations. (1) Naïve was very slow. Even though it applies Split rather than finding the max flow for every pair of nodes, so already optimizes by Theorem 4.14, it took 6.8 hours on only 20% data and took more than 10 hours on 40% data. (2) INDEX improved Naïve by two orders of magnitude just because the index simplifies finding neighborhood v-cliques; however, it still took more than 10 hours on 80% data. (3) SINDEX improved INDEX by 41% on 60% data as it reduces the size of the inverted index by 64%. (4) UNION improved SINDEX by 47% on 60% data; however, it also took more than 10 hours on 80% data. (5) Pivot improved UNION significantly; it finished in 2.2 minutes on the whole data set so further reduced execution time by at least three orders of magnitude, showing importance of splitting.

Stage II: After Stage I we have .7M pivots and 17.3M remaining records. It spent 6.4 hours for Stage II: 1.7 hours for blocking and 4.7 hours for clustering. The long time for clustering is because of the huge number of blocks. There are 1.4M blocks with multiple elements (a pivot is counted as one element), with a maximum size of 22.5K and an average of 4.2. On only 35 blocks clustering took more than 1 minute and the maximum is 2.5 minutes, but for 99.6% blocks the size is less than 100 and ClUSTER took less than 60 ms. The average time spent on each block is only 9.6 ms.

VII. Conclusions

In this paper we studied how to link records to identify groups. We proposed a robust two-stage algorithm that is shown to be empirically scalable and accurate over two real-world data sets. Future work includes extending our techniques to find overlapping groups, and applying the two-stage framework in other contexts where tolerance to value diversity is critical, and erroneous data is prevalent.

REFERENCES