Combining Large Datasets of Patents and Trademarks

Grid Thoma
Computer Science Division, School of Science & Technology
University of Camerino

14th Italian STATA User Annual Meeting
Florence, 16 Nov 2017

Motivations

- Where do innovators come from?
  - location, industry, cohort, size, listing, VC, …
- How to appraise correctly IP counts at the patentee's portfolio level?
  - Patents, trademarks, and designs
  - EPO, WIPO, USPTO, … , families of priority links
  - Citations / self-citations
- The problem of harmonization of entity names

Different spellings/misspellings

MINNESOTA MINING AND MANUFACTURING COPANY
MINNESOTA MINING AND MANUFACTURING COPMANY
MINNESOTA MINING AND MANUFACTURING CORP
…
BSH BOSCH UND SIEMENS AKTIENGESELLSCHAFT
BSH BOSCH UND SIEMENS AKTINGESELLSCHAFT
BSH BOSCH UND SIEMENS HANSGERAETE GMBH
BSH BOSCH UND SIEMENS HAUS-GERAETE GMBH
BSH BOSCH UND SIEMENS HAUSERATE GMBH

Variations in naming conventions

MINNESOTA MINING & MFG CO

3M CORP

MINNESOTA & MINING MANUFACTURING

...

INTERNATIONAL BUSINESS MACHINES – IBM

IBM CORP. (INTERNATIONAL BUSINESS MACHINES)

IBM CORPORATION (INTERNATIONAL BUSINESS MACHINES)

Assignment to aggregate entities (ownership issues)

Subsidiaries with parent MINNESOTA MINING & MFG CO:

ADHESIVE TECHNOLOGIES INC
AVI INC
D L AULD CPY
DORRAN PHOTONICS INCORPORATED
EOTEC CORPORATION
NATIONAL ADVERTISING CPY
RIKER LABORATORIES INC
TRIM LINE INC

Sources

- NBER Patent Data Project *(harmonized entity names)* sites.google.com/site/patentdataproject
- USPTO's data disclosure initiative *(in STATA files)* www.uspto.gov/economics
- Magerman *et al.* (2006). Data production methods for harmonized patent statistics: Patentee name standardization. KU Leuven FETEW MSI.
- Thoma *et al*. (2010). Harmonizing and combining large datasets – an application to firm-level patent and accounting data. NBER WP # 15851.

Agenda

- Background
- Dataset
- Software creation and results
- Quality checks

Agenda

- **Background**
- Dataset
- Software creation and results
- Quality checks

Dictionary based approach

- Large collections of entity names, serving as examples for a specific entity class
- Exact matching of dictionary entries OR
- … "fuzzify" the dictionary by (automatically) generating typical spelling variants for every entry
- The problem of recall rate

(*e.g. ANSI / UNICODE*)

Articulation of a dictionary

- Every known variation of an entity name
- Harmonized to one agreed standard name

Applicants

Variation name	Standard name
AKTIENGESELLSCHAFT VOLKSWAGEN	VOLKSWAGEN AG
FUORUKUSUAAGENUERUKU AG	VOLKSWAGEN AG
FUORUKUSUWAAGEN AG	VOLKSWAGEN AG
V O L K S W A G E N AKTIENGESE	VOLKSWAGEN AG
V W AG	VOLKSWAGEN AG
VOLKSWAGEN	VOLKSWAGEN AG
VOLKSWAGEN A G	VOLKSWAGEN AG
VOLKSWAGEN AG	VOLKSWAGEN AG
VOLKSWAGEN AG VW	VOLKSWAGEN AG
VOLKSWAGEN AKTIENGESELLSCHAFT	VOLKSWAGEN AG
VW	VOLKSWAGEN AG
VW AG	VOLKSWAGEN AG
VW WOLFSBURG	VOLKSWAGEN AG
WOLFSBURG VW	VOLKSWAGEN AG
BRASI S A VOLKSWAGEN DO	VOLKSWAGEN BRASIL
BRASIL S A VOLKSWAGEN DO	VOLKSWAGEN BRASIL
BRASIL SA VOLKSWAGEN	VOLKSWAGEN BRASIL

Existing dictionaries of patenting entity names

- USPTO / EPO standard patentee codes
- DERWENT patentee codes
- NBER Patent Data Project (*file: patassg.dta*) sites.google.com/site/patentdataproject
- Harmonization procedure to build a dictionary (Magerman *et al.* 2006)

Magerman *et al.* (2006)'s procedure

1. Character cleaning
2. Punctuation cleaning
3. Legal form indication treatment
4. Spelling variation harmonization
5. Umlaut harmonization
6. Common company name removal
7. Creation of a unified list of entity names

Rule-based approach

- Definition of rules to compare the similarity of names (Thoma *et al.* 2010)
- Initially, hand-crafted rules to describe the composition of named entities and their context
- Some core words and components of words used to extract candidates for more complex names
- … OR viceversa

Approximate string matching algorithms (1)

- **Edit distance**: the minimum number of operations to switch from one word to another
  - Typically used to account for spelling variations
  - Similarity of two strings *x* and *y* of length n_x and n_y calculated as

$$1-d/N$$

where *1* is the maximum similarity;

d is the distance between *x* and *y*;

$N=max\{n_x, n_y\}$.

Edit distance: examples

1. HILLE & MUELLER GMBH & CO./
   HILLE & MULLER GMBH & CO KG /
   HILLE & MÜLLER GMBH & CO KG

2. AB ELECTRONIK GMBH/
   AB ELEKTRONIK GMBH

3. BHLER AG / BAYER AG

Approximate string matching algorithms (2)

- **Jaccard Similarity measure**: number of unique common tokens of two strings divided by the number of tokens in the union

$$J = \frac{T_1 \cap T_2}{T_1 \cup T_2}$$

Approximate string matching algorithms (2)

- **Jaccard Similarity measure**: number of unique common tokens of two strings divided by the number of tokens in the union

$$J = \frac{T_1 \cap T_2}{T_1 \cup T_2}$$

- **Computationally Easy *J* Similarity Measure**:

$$J \cong 2\frac{T_1 \cap T_2}{T_1 + T_2}$$

Jaccard similarity: examples

1. AAE HOLDING / AAE TECHNOLOGY INTERNATIONAL
2. JAPAN AS REPRESENTED BY THE PRESIDENT OF THE UNIVERSITY OF TOKYO /PRESIDENT OF TOKYO UNIVERSITY
3. AAE HOLDING / AGRIPA HOLDING
4. VBH DEUTSCHLAND GMBH / IBM DEUTSCHLAND GMBH

Approximate matching algorithms (3)

Weighted Jaccard Similarity Measure

- Inversely weighted by the frequency n_i of a given token *i* across different entity names

$$J^w(X,Y) = \frac{2\sum_{k|x_k \in X \cap Y} w_k}{\sum_{i|x_i \in X} w_i + \sum_{j|y_j \in Y} w_j}$$

where

$$w_i = \frac{1}{log(n_i) + 1}$$

Agenda

- Background
- **Dataset**
- Software creation and results
- Quality checks

Patent and trademark datasets

- Patenting entity names at the USPTO
  - Reference dictionary *(NBER Patent Data Project)*
  - A unique ID code for a patentee *(file: patassg.dta*)
- Trademarking entity names at the USPTO
  - www.uspto.gov/economics *(file: owner.dta)*
- Time coverage
  - Patents: 1976-2006; Trademarks: 1977-2015
- Focus: US business organizations
  - 117,443 unique ID codes from the reference dictionary
  - 3,462,601 (unharmonized) trademarking entity names
- Entity name matching executed within state level

Harmonization of address information

- Only state & city info in patent records
- Full address info for trademarks
  - 5 digit zip codes in 98.5% of the US addresses
- Harmonization of city names
  - Removing numbers & non standard chars
- Geocoding based on geonames.usgs.gov
- Edit distance / Soundex for matching city names

Agenda

- Background
- Dataset
- **Software creation and results**
- Quality checks

STATA implementation (1)

- An augmented harmonization procedure to create a dictionary for the trademarking entity names (Thoma *et al.* 2010)
- J^W similarity measure for the matching of the patenting & trademarking entity name dictionaries
- Location information to reduce false positives and false negatives
- Manual inspection to improve accuracy and matching rate
- Improvement of dictionary use through priority links

STATA implementation (2)

1. Reshape entity names as tokens in long format
2. Remove non standard chars & numbers
3. Drop single char tokens
4. Pool tokens to create a dictionary of tokens
5. Inflate the dictionary with tokens from patent titles / wordmarks (improving statistical weights)
6. Drop stop words (frequent/non discriminating)
7. Compute the defined statistical weight of a token

STATA implementation (3)

8. Merge files based on tokens and state level codes of an entity name
9. Collapse the tokens' statistical weights to compute the J^W measure's numerator of a matched pair
10. Compute the J^W measure, including the denominator
11. Sort matched pairs based on the J^W measure, selecting the best match

Figure 1: Share of US business patentees matched with trademarks
(Notes: States with 1000+ patentees; Source: USPTO)

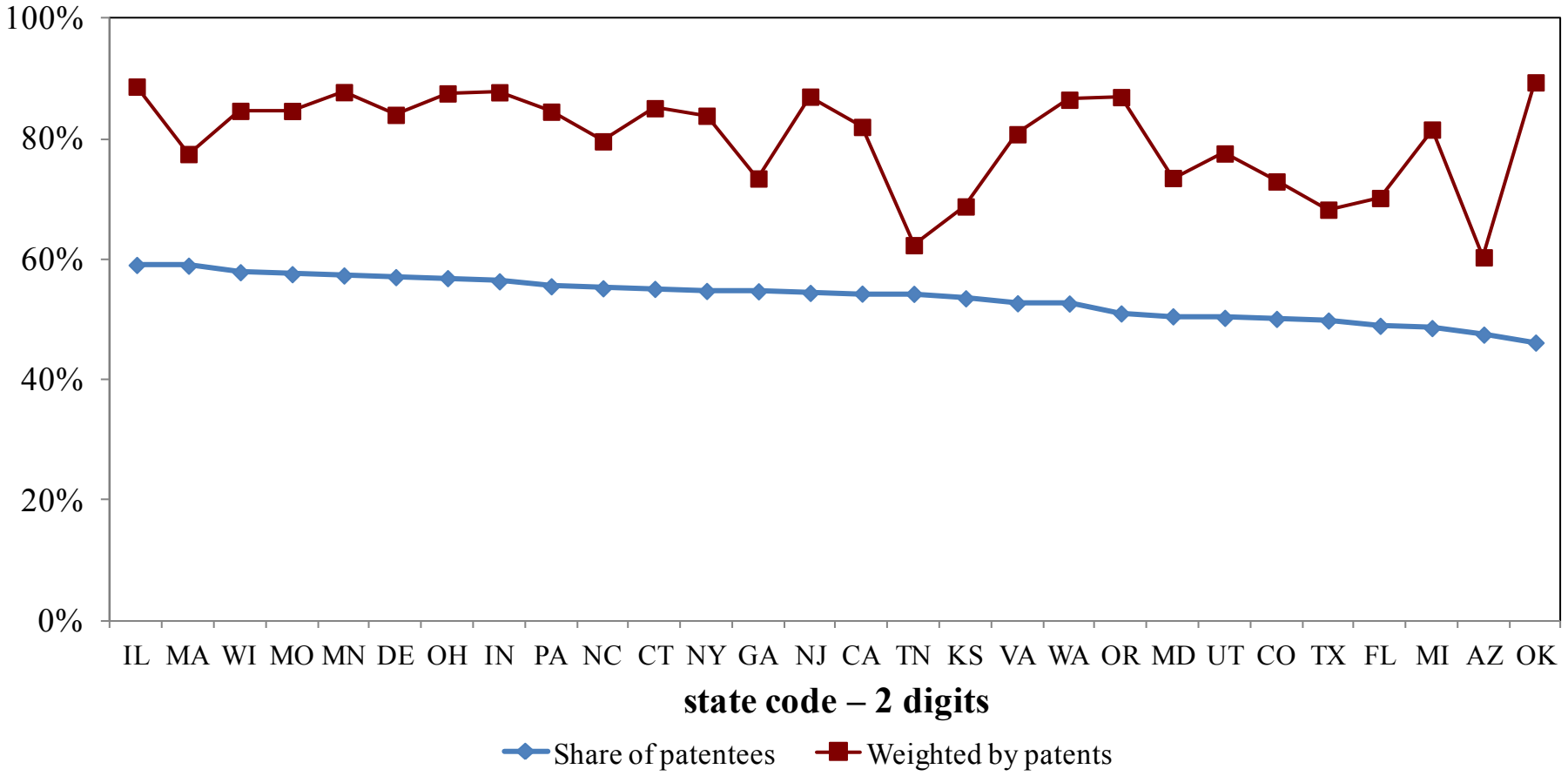
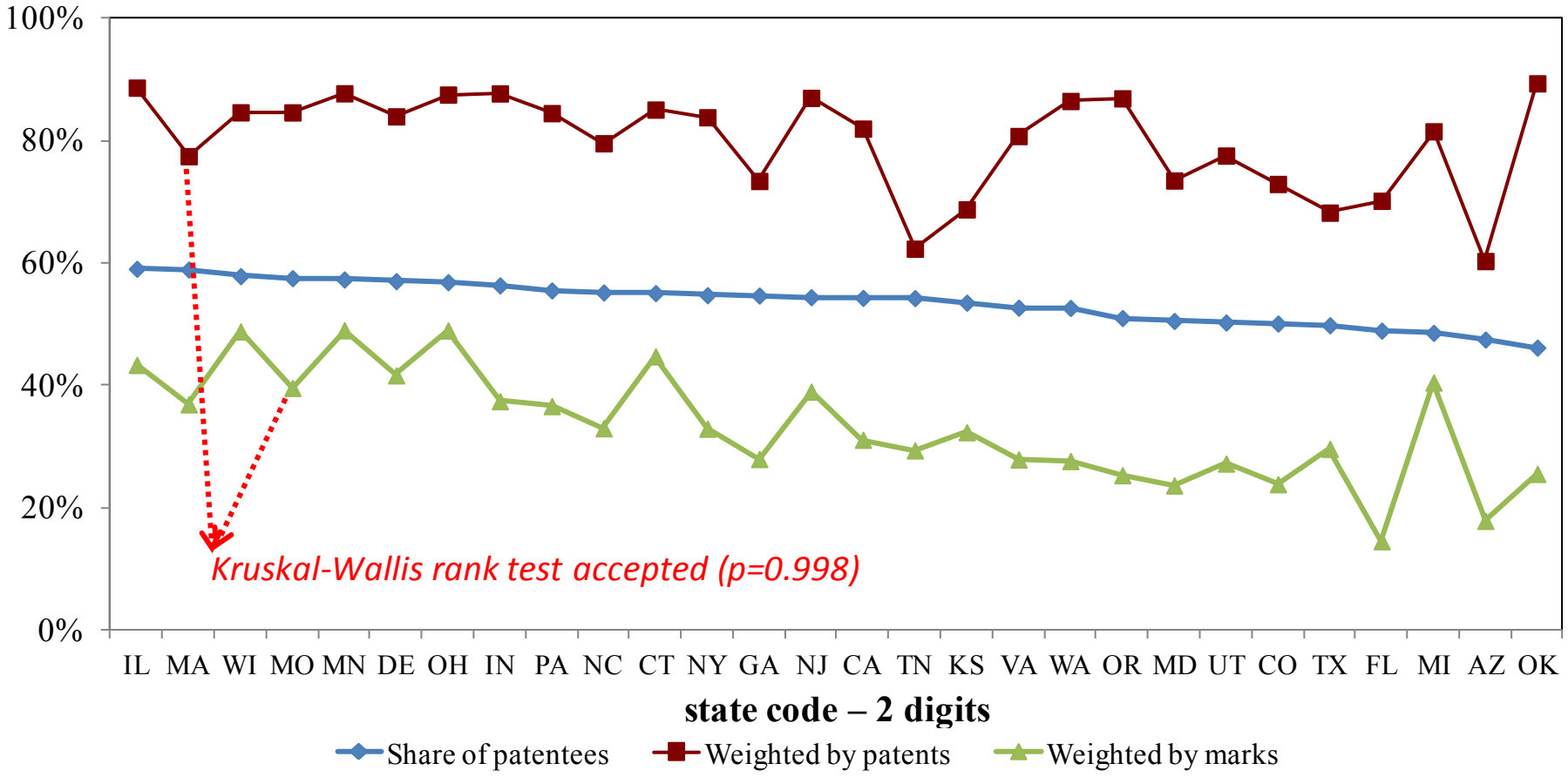


Figure 1: Share of US business patentees matched with trademarks
(Notes: States with 1000+ patentees; Source: USPTO)



Agenda

- Background
- Dataset
- Software creation and results
- **Quality checks**

Selection of the best match

- Below a certain threshold of J^w, select the best match with the highest J^w
- Define a goodness index (*matching score*) of a matched pair using J^w & address information (state–city correspondence)
- Manual inspection in order to define the appropriate thresholds of the *matching score*
- Select the best match with the lowest *matching score*

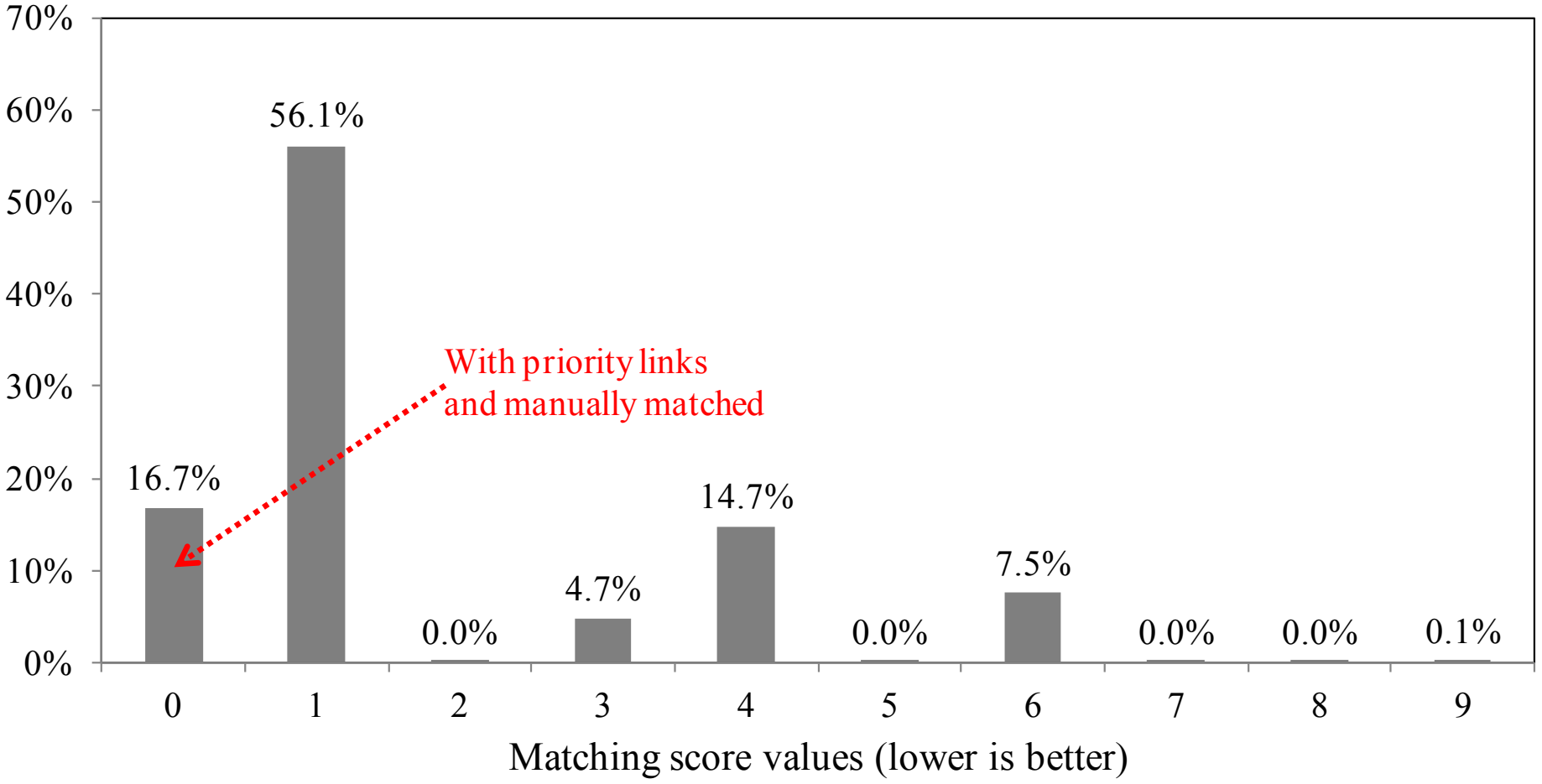
Selection of the best match through the matching score

For each matched name a mutually exclusive goodness score is given from 1-9, where:

J^w Similarity Measure	Same location	Unknown location	Different location
$J^w \geq 67\%$	**1**	**2**	**3**
$57 \leq J^w < 67\%$	**4**	**5**	**6**
$47 \leq J^w < 57\%$	**5**	**8**	**9**

Thresholds defined through manual scrutiny

Figure 2 Distribution of the matching score of the matched names: US business patentees matched to the trademarking entity names



Improvement of dictionary usage through priority links

- Priority links in patents and trademarks
- Potential limitations
  - Copatentees of a patent/trademark
  - Entity name changes (synonymies)
  - Subsidiaries
  - Distinct entity names
  - Entity address changes

Harmonization tasks of entity names through priority links

- Focus on the trademarking entity names
- Retrieve forward/backward priority links
- Consolidate links to create self containing families of priorities
- Manual scrutiny in merging families with standard entity names
- In the overall dataset, propagate standard entity names using perfect name matching, and having the same zip code

Diagnostics: resolving duplicate matching candidates (potential)

- **The earliest patenting entity**
- Technological-market affinity
- Name changes over time
- Ownership structure of companies

Figure 3. Time lag of the first trademark since year of the first patent

(Notes: US business patentees active with patenting & trademarking during 1981–2003;
Source: USPTO)