# SearchEngine - a universal linkage tool

Thorsten Doherr

https://github.com/ThorstenDoherr/searchengine

# Main application: firm data linkage

- Patent portfolio on firm level
  - Technological orientation, change over time (e.g. adaption “Industry 4.0”)
  - Citation networks among firms
  - Proxy for innovation
  - …
- Publication portfolio on firm level
  - Substitution of patents (strategical publications)
  - Cooperation with research institutions or universities
  - …
- Subsidy data on firm level
  - Evaluation of subsidy programs regarding firm parameters, patents, publications, cooperation, …
- Person data on firm level
  - Relating performance of startups with the educational background of the founder teams
  - Impact of inventor mobility among firms
  - …
- …

# Other applications

- Disambiguation: identifying similar entries within a database
  - Removing duplicate firm entries
  - Clustering of applicants in patent data
  - Clustering of titles or abstracts
  - Inventor name disambiguation
  - …
- Removing overlap
  - Cleaning a control group from treated entries, e.g. subsidized firms
  - Removing duplicates in a sample extension for the next wave of a panel survey
  - …

# The problem: no mutual firm-ID

- Provider with dedicated firm focus
  - „Clean“ data is business model or statutorily necessary
  - Proprietary provider IDs but also statutory numbers, i.e. VAT-ID
  - Active avoidance of ambiguous entries by manual and/or automatic examination
- Provider without dedicated firm focus
  - Administer other contents (e.g. patents, publications, subsidies) degrading firms to mere appendages
  - Firms don’t have separate IDs. They are an attachment to the main context key usually identifying a document.
  - The authentic reproduction of the document is the focus which entails the diligent transfer of misspellings and inconsistent wording (and introduction of new ones)
- Linkage: firm name and available address
  - Different wording: abbreviations, word positioning, redundancy
  - Different context: firms vs. departments vs. branches
  - Incompatible fields: one sweeping address field vs. distinct fields for street, zip, city
  - Misspellings

# Harmonization

## Reducing variation by applying normative functions

- Character transformations
  - Everything to UPPER or lower case
  - Replacing umlaut letters, “âcçènts” and other language specific characters by a simplified ASCII representation, e.g. ä → ae, ß → ss
  - Removing non-alphanumeric characters
- Semantic transformations
  - Context specific transformation of/to common abbreviations, e.g. LTD → LIMITED
  - Removing duplicate and filler words, e.g. “and”, “of”, “the”
    Some contexts, like legal forms, may also be considered redundant
- Destructive transformations
  - Sorting of words within a field
  - Truncation of words
  - Limitation of word count
  - Phonetic methods like Soundex, Metaphone, Cologne Phonetics, n-grams
  - …

# Limitations of harmonization

## Matching datasets after harmonization by simple joins

- Identification of filler words based on common knowledge is never exhausting (besides the person conducting it)
- Additional words or missing words, which are not fillers, lead to different harmonized entries for the same entity
  - This issue is exacerbated if both datasets are contextually different, i.e. titles vs. abstracts, firm name vs. firm name plus department
- Destruction of information improves robustness against misspellings but increases the risk of false positives significantly
- A match is a match: no intrinsic quality measure, always 100%

  → Matches are always commutative and transitive
  - Commutativity: if A = B then B = A
  - Transitivity: if A → B and B → C then A → C
    Enforcing transitivity imposes restrictions on the match

# The challenge of harmonization

KATO HATSUJO KAISHA LTD
KATOU HATSUJO KAISHA LTD
KATO HATSUJO KAISHA LIMITED
KATO HATSUJO COMPANY LIMITED
KATO HATSUJI KAISHA LTD
KATO HATSUJO COMPANY LTD
KATO HATSUJO KABUSHIKI KAISHA

Do you know all Japanese legal forms?

Is “SECRETARY” a filler word?
Reordering of words might help!

Good luck finding a harmonization for the last one!

THE COMMONWEALTH OF AUSTRALIA DEPARTMENT OF DEFENCE
COMMONWEALTH OF AUSTRALIA DEPARTMENT OF DEFENCE
THE COMMONWEALTH OF AUSTRALIA CO THE SECRETARY DEPARTMENT OF DEFENCE
THE COMMONWEALTH OF AUSTRALIAA DEPARTMENT OF DEFENCE
CO THE SECRETARY COMONWEALTH OF AUSTRALIA DEPARTMENT OF DEFENCE
THE COMMONWEALTH OF AUSTRALIA THE SECRETARY DEPARTMENT OF DEFENCE

HUHTAMAKI FORCHHEIM ZWEIGNIEDERLASSUNG DER HUHTAMAKI DEUTSCHLANG GMBH CO KG
HUHTAMAKI FORCHEIM ZWEIGNIEDERLASSUNG DER HUHTAMAKI DEUTSCHLAND GMBH CO KG
4P FOLIE FORCHHEIM ZWEIGNIEDERLASSUNG DER HUHTAMAKI VAN LEER DEUTSCHLAND GMBH CO KG
4P FOLIE FORCHHEIM GMBH
HUHTAMAKI FORCHHEIM ZWEIGNIEDERLASSUNG DER HUHTAMAKI DEUTSCHLAND GMBH CO KG
4P FOLIE FORCHHEIM ZWEIGNIEDERLASSUNG DER VAN LEER DEUTSCHLAND GMBH CO KG

Source: USPTO patent applicants

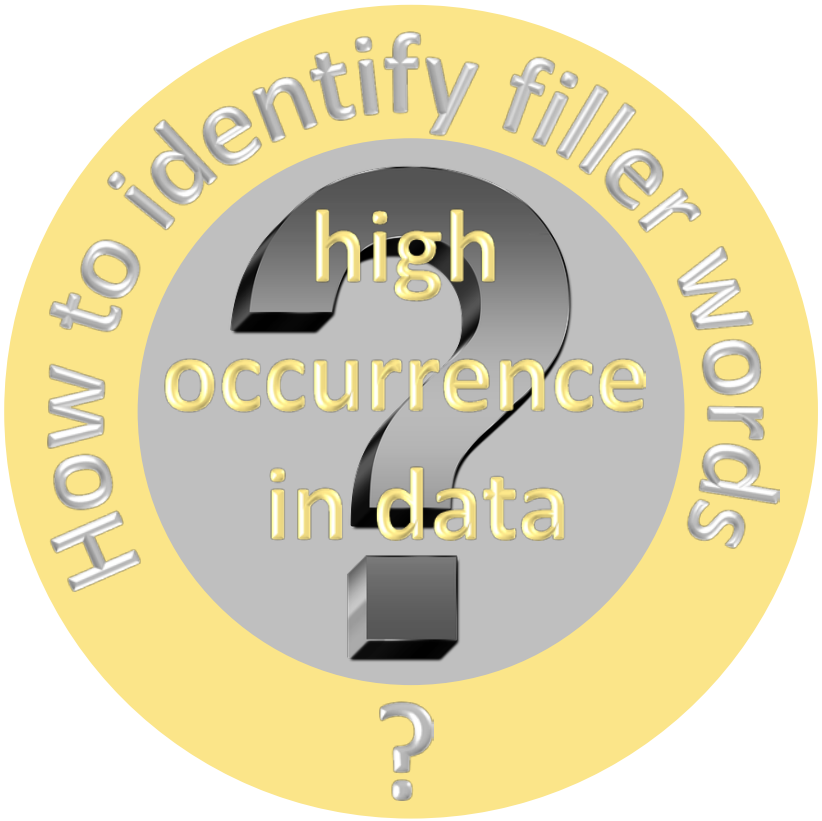
# In the eyes of the SearchEngine

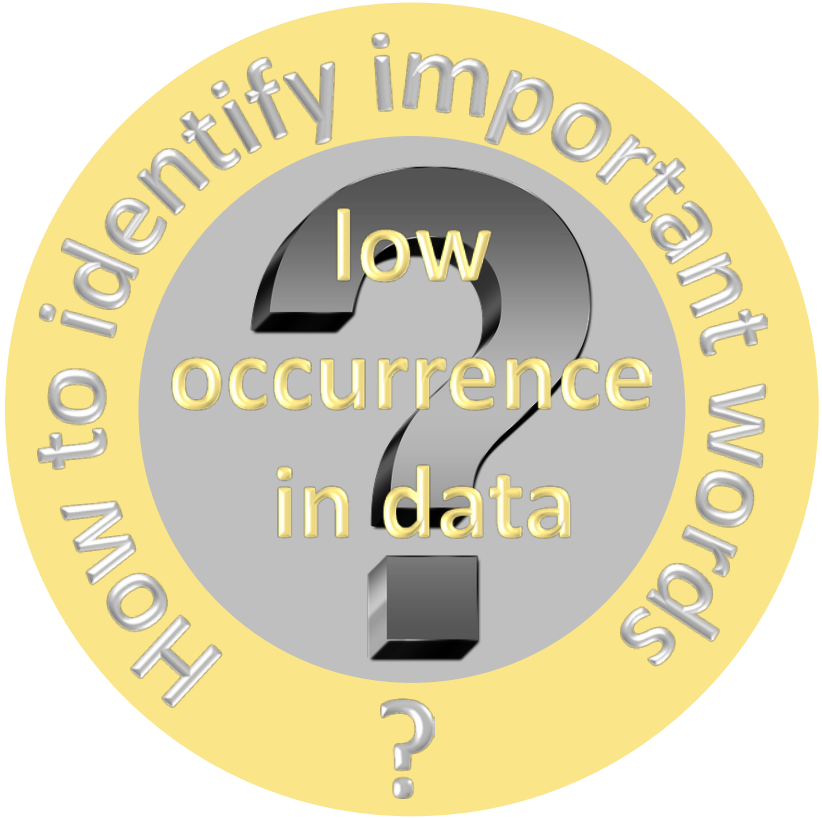
KATO HATSUJO
KATOU HATSUJO
KATO HATSUJO
KATO HATSUJO
KATO HATSUJI
KATO HATSUJO
KATO HATSUJO

COMMONWEALTH AUSTRALIA DEPARTMENT DEFENCE
COMMONWEALTH AUSTRALIA DEPARTMENT DEFENCE
COMMONWEALTH AUSTRALIA SECRETARY DEPARTMENT DEFENCE
COMMONWEALTH AUSTRALIAA
COMONWEALTH DEFENCE
COMMONWEALTH AUSTRALIA SECRETARY DEPARTMENT DEFENCE

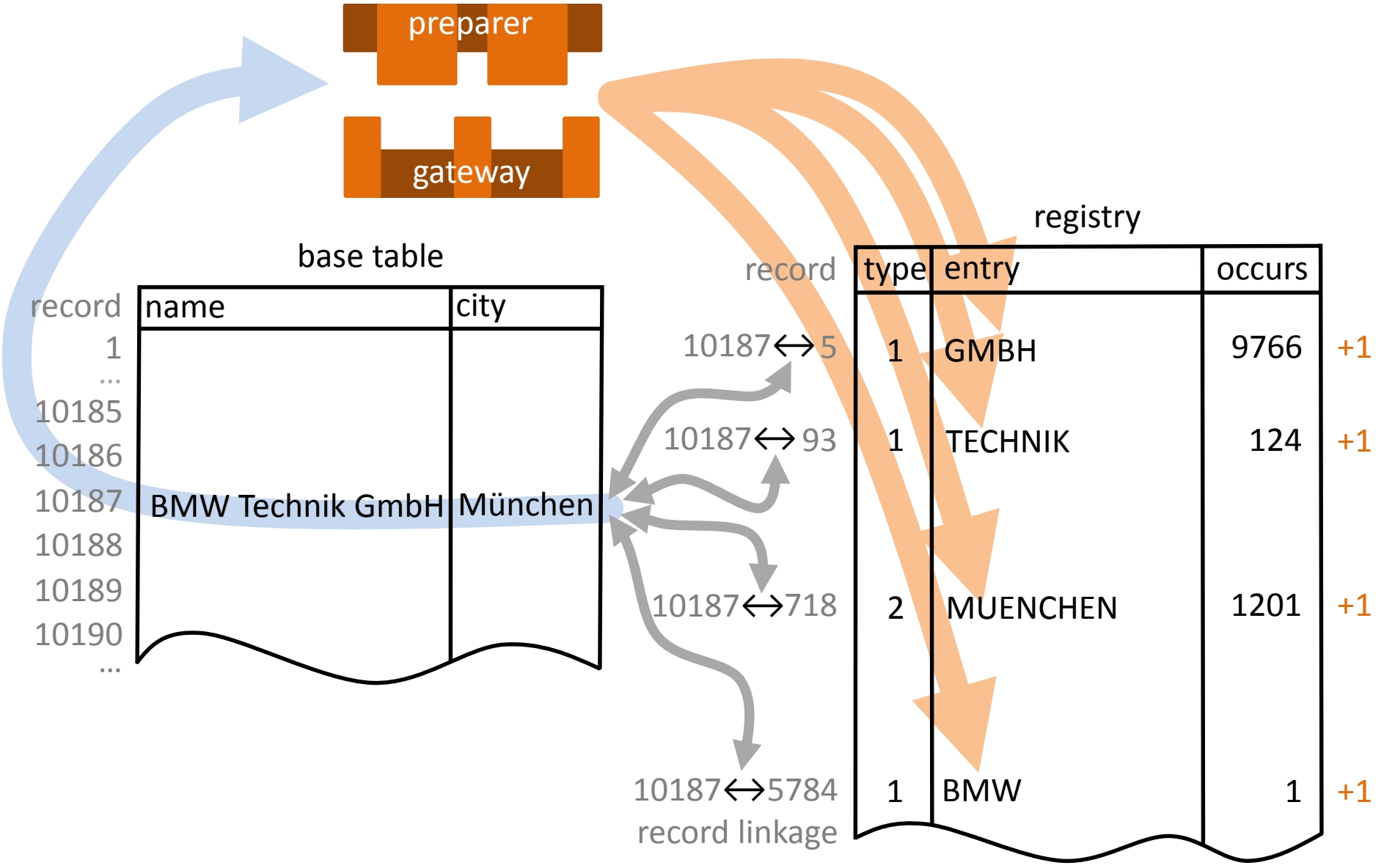
HUHTAMAKI FORCHHEIM ZWEIGNIEDERLASSUNG HUHTAMAKI DEUTSCHLANG
HUHTAMAKI FORCHEIM ZWEIGNIEDERLASSUNG HUHTAMAKI
4P FOLIE FORCHHEIM ZWEIGNIEDERLASSUNG HUHTAMAKI LEER
4P FOLIE FORCHHEIM
HUHTAMAKI FORCHHEIM ZWEIGNIEDERLASSUNG HUHTAMAKI
4P FOLIE FORCHHEIM ZWEIGNIEDERLASSUNG LEER

# Two sides of the same coin



# Components of the SearchEngine



# SearchEngine components

- Preparer gateway
  - Obligatory base line harmonization: character transformation
  - Optional harmonization by so called preparer: semantic and destructive transformation
  - Tokenizing
- Registry
  - Counts words of the base table
  - Base table is the largest table or the most noisy one providing the candidates of a search
  - Search tables are all other tables, usually smaller ones or with less noise because noise dilutes the context causing more harm in a search term than in the candidates
  - Relevant search fields in the base table are represented by chapters in the registry, called search types. Different combinations of search fields and preparer create new search types.
  - Generates summary statistics per search type like maximum and average occurrence. The latter is used to impute unknown words.
- Record linkage
  - Fast access to all registry entries belonging to a given base table record
  - Fast access to all base table records sharing a specific registry entry

# Basic heuristic

| name | 1 | street | 2 | zip | 3 | city | 4 |
|---|---|---|---|---|---|---|---|
| BMW FORSCHUNG UND TECHNIK GMBH | | HANAUER STRASSE 46 | | 80992 | | MUENCHEN | |

| type | entry | occurs | $IP = \frac{1}{occurs}$ | $share = \frac{IP}{\sum IP}$ | $rIP = share * weight$ |
|---|---|---|---|---|---|
| … | … | … | | | |
| 1 | BMW | 552 | 0.0018116 | 76.863% | 53.804% |
| 1 | FORSCHUNG | 1980 | 0.0005051 | 21.428% | 15.000% |
| 1 | TECHNIK | 25552 | 0.0000391 | 1.660% | 1.162% |
| 1 | UND | 1190073 | 0.0000008 | 0.036% | 0.025% |
| 1 | GMBH | 3353864 | 0.0000003 | 0.013% | 0.009% |
| … | … | … | ∑ 0.0023569 | ∑ 100.00% | ∑ 70.000% |
| 2 | HANAUER | 6851 | 0.0001460 | 90.511% | 9.051% |
| 2 | 46 | 65931 | 0.0000152 | 9.494% | 0.941% |
| 2 | STRASSE | 7410645 | 0.0000001 | 0.084% | 0.008% |
| … | … | … | ∑ 0.0001613 | ∑ 100.000% | ∑ 10% |
| 3 | 80992 | 3905 | 0.0002561 | 100.000% | 10% |
| … | … | … | ∑ 0.0002561 | ∑ 100.000% | ∑ 10% |
| 4 | MUENCHEN | 316874 | 0.0000032 | 100.000% | 10% |
| … | … | … | ∑ 0.0000032 | ∑ 100.000% | ∑ 10% |

# Relative Identity

| name | | | | | 1 | street | | | 2 | zip | 3 | city | 4 | $\sum rIP$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **BMW** | **FORSCHUNG** | **UND** | **TECHNIK** | **GMBH** | | **HANAUER** | **STRASSE** | **46** | | **80992** | | **MUENCHEN** | | |
| 53.8 | 15.0 | 0.0 | 1.2 | 0.0 | | 9.1 | 0.0 | 0.9 | | 10 | | 10 | | 100.00% |

| | | | | |
|---|---|---|---|---|
| **BMW FORSCHUNG UND TECHNIK GMBH** | **HANAUER STRASSE 46** | **80992** | **MUENCHEN** | 100.00% |
| **BMW FORSCHUNG U TECHNIK GMBH** | **HANAUER STRASSE 46** | **80992** | **MUENCHEN** | 99.98% |
| **BMW TECHNIK UND SERVICE GMBH** | **HANAUER STRASSE 48** | **80992** | **MUENCHEN** | 84.06% |
| **BMW STIFTUNG HERBERT QUANDT** | **HANAUER STRASSE 46** | **80992** | **MUENCHEN** | 83.80% |
| **BMW MAENNERCHOR MUENCHEN EV** | **DACHAUER STRASSE 371** | **80992** | **MUENCHEN** | 73.81% |

Threshold: 70%

# Not commutative $(a \rightarrow b \neq b \rightarrow a)$

| name | | | 1 | street | | 2 | zip | 3 | city | 4 | $\sum rIP$ |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **BMW** | **MAENNERCHOR** | **MUENCHEN** | **EV** | **DACHAUER** | **STRASSE** | **371** | **80992** | | **MUENCHEN** | | |
| 18.0 | 51.2 | 0.8 | 0.0 | 1.2 | 0.0 | 8.8 | 10 | | 10 | | 100.00% |

| | | | | |
|---|---|---|---|---|
| **BMW MAENNERCHOR MUENCHEN EV** | **DACHAUER STRASSE 371** | **80992** | **MUENCHEN** | 100.00% |
| **MANNERCHOR** RIESENFELD EV | ABBACH **STRASSE** 27 A | **80992** | **MUENCHEN** | 71.23% |

Threshold: 70%

# Heuristic: relative identification potential *rIP*

$$IP(w) = occ\big(w, st(w)\big)^{-1}$$

$$rIP(w) = weight(st(w)) \left( \frac{IP(w)}{\sum_{v \in S} \begin{cases} IP(v) & | \; st(w) = st(v) \\ 0 & | \; st(w) \neq st(v) \end{cases}} \right)$$

$S$ is the search term (set of searched words)
$w$ is a word of the search term: $w \in S$
$st(w)$ returns the search type of word $w$
$occ(w, t)$ retrieves the occurrence of word $w$ for search type $t$ from the registry
$weight(t)$ returns the weight of search type $t$

# Heuristic: absolute identification potential score

An arbitrary measure of the absolute identification potential of a search term to rank candidates by the score of their search term

$$score(S) = \sum_{w \in S} \frac{weight\big(st(w)\big)}{occ\big(w, st(w)\big)}$$

$S$ is the search term (set of searched words)
$w$ is a word of the search term: $w \in S$
$st(w)$ returns the search type of word $w$
$occ(w, t)$ retrieves the occurrence of word $w$ for search type $t$ from the registry
$weight(t)$ returns the weight of search type $t$

# General data preparation

- Assessment of the data to determine search direction (search → base)
  - Base table: larger and/or noisier, e.g. firm name sporadically contains departments
    Defines the heuristic → surplus words do not affect results
  - Search table: smaller and/or cleaner, e.g. firm focused data supplier, less noise
    Provides the search terms → surplus words affect the results
  - Using focused data as base table returns cleaner results with less false positives
    - Patent applicants → focused firm data: use only the best candidate
    - Focused firm data → patent applicants: take all candidates to catch all variants
- Adjustment of data and contexts
  - If possible: separation of contexts, e.g. address → street, zip, city
    More fields → more options for weight distribution → higher precision
  - If not: combination of fields to establish parity between search and base table structure
    Context overlaps lead to lower precision (more false positives)
  - Individual adjustments, e.g. transformation of Chinese Unicode characters into their hexadecimal representation (**赤** → [2F9A])
- Aggregation by relevant fields, e.g. firm name, street, city, zip
  - Elimination of identical entries: every record is unique in terms of search fields
  - Linkage of aggregated data to original data, e.g. in Stata:

```
egen long unikey = group(name street city zip), missing
save original
duplicates drop unikey, force
save search_or_base
```

# Additional features

- Misspellings

  These aberrations have a large impact on the heuristic because they usually are rare enough to become dominant. We discuss methods to tame them.

- Smoothing/Accentuating

  Smoothing flattens the weight distribution per search type to provide a more balanced search behavior. Accentuating does the opposite.

- Feedback

  The basic heuristic ignores surplus words not corresponding with the search term in the candidates. Feedback introduces gradual commutativity.

- Cutoff and activation

  Retrieving a copious amount of candidates is an indicator for a weak search term prone to false positives. Cutoff and activation is a simple method to handle these hopeless cases.

- Historical data and variants

  Variants or identifiable historical versions of an entity in the data improve the success rate but also increase redundancy, which can be suppressed on the entity level.

- Incremental search strategies

  The “one search fits all” does not exists. Multiple search strategies have to be applied to a search task to collect all candidates without swamping the result table with false positives.

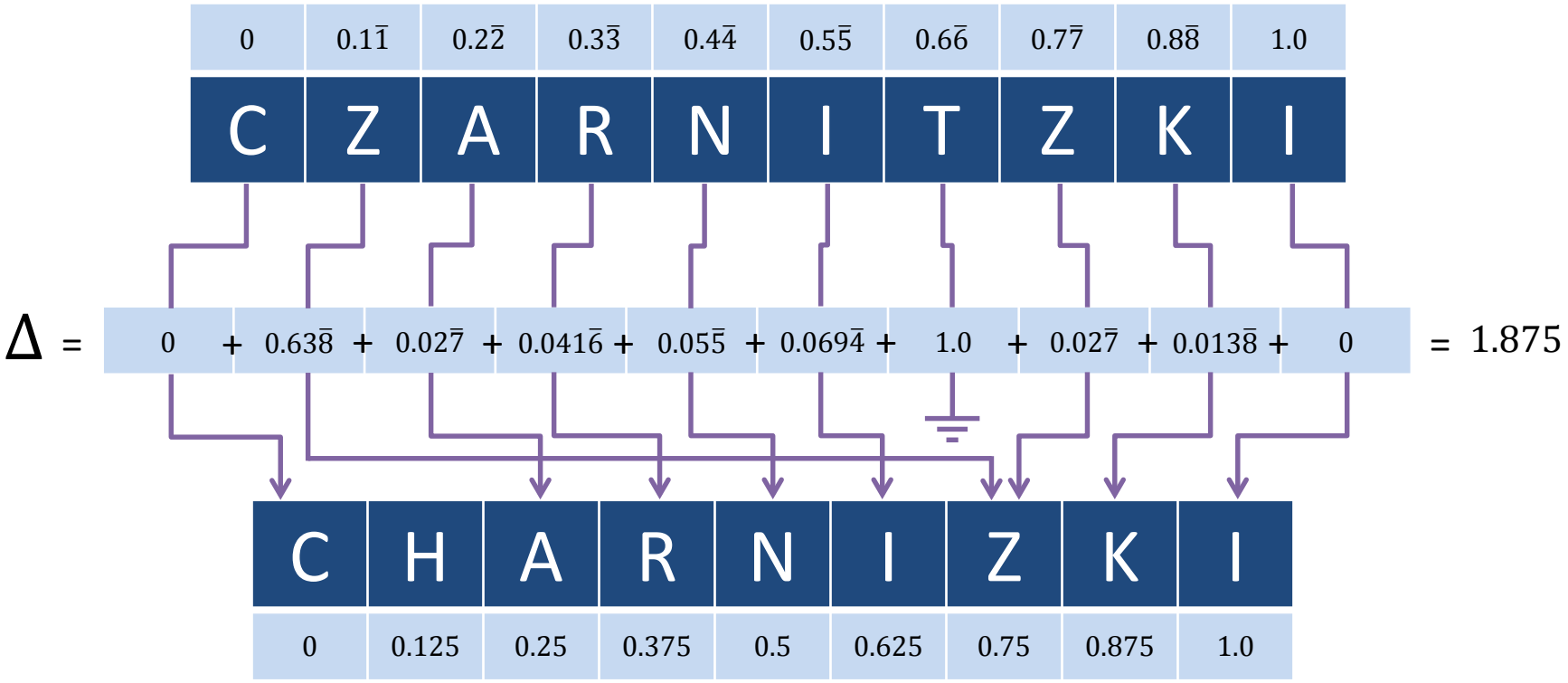
- Disambiguation

  When search and base table are the same.

# Misspellings

- Search types based on destructive preparer
  - Soundex(“CZARNITZKI”) = “C26532”
  - Metaphone(“CZARNITZKI”) = “KSRNTSK”
  - Gram(4, “CZARNITZKI”) = “CZAR ZARN ARNI RNIT NITZ TZKI”
  - Gram(3, “CZARNITZKI”) = “CZA ZAR ARI RNI NIT TZK ZKI”
- Destroying information for the sake of higher robustness to misspellings leads to a strong tendency for false positives
- Subsequent evaluation of candidates
  - Based on original content (before application of destructive preparer)
    → restoration of destroyed information
  - String distance function compares search fields with candidate fields word by word looking for the respective best fit
    → independent from word order (like the basic heuristic)
  - Returns a percentage as a measurement of the similarity
    → simple integration by replacing the $rIP$ of the respective search type

# Least Relative Char Position Delta (LRCPD)




$$lrcpd(word1, word2) = 1 - \frac{\Delta(word1, word2)}{len(word1)} = 1 - \frac{1.875}{10} = 0.8125$$

# Phonetics vs. fragmentation

- Phonetic methods
  - Efficient in terms of computational resources
  - High similarity of words sharing the same code
  - Encoding reflects particularities of the source language
    - English: metaphone, soundex
    - German: cologne (Kölner Phonetik)
  - Not suited for international data sources
  - Sensitive to separation or concatenation of words
- Fragmentation into n-grams
  - High strain on computational resources
  - Retrieved candidates can be anagrams of the search term
  - Robust to truncation or concatenation of words
  - Language independent

# Smoothing/Accentuating

## Shift from nuanced weights to uniformly or distinctively weighted words

- Applicable per search type
- Smoothing
  - Enforcing a more conservative search behavior requiring a larger share of words to match
  - When the basic heuristic is inappropriate
    - Example: house numbers in street addresses
      High numbers have always a higher $IP$ than low numbers because every street has a number 1, but only few have the number 999
    - Example: person names
      First names usually have less variation than last names, but still can not be considered filler words
    - Search types based on n-grams can still be vulnerable if a misspelling causes exotic grams
- Accentuating
  - Strengthens dominant words to counter noise in search terms (caution: may backfire)
- Methods
  - Offset
  - Logarithmic inverse word frequency ratio
  - Softmax

# Smoothing methods

## Offset

$$IP(w) = max\big(occ\big(w, st(w)\big) + off\big(st(w)\big), 1\big)^{-1}$$

## Logarithmic inverse word frequency ratio

$$IP(w) = max\left(ln\left(\frac{maxocc\big(st(w)\big)}{occ\big(w, st(w)\big)}\right), 1\right)$$

## Combined

$$IP(w) = max\left(ln\left(\frac{max\big(maxocc\big(st(w)\big) + off\big(st(w)\big), 1\big)}{max\big(occ\big(w, st(w)\big) + off\big(st(w)\big), 1\big)}\right), 1\right)$$

$st(w)$ returns the search type of word $w$
$occ(w, t)$ retrieves the occurrence of word $w$ for search type $t$ from the registry
$off(t)$ returns the offset for search type $t$
$maxocc(t)$ returns the maximum occurrence for search type $t$

# Softmax

Smooths or accentuates the distribution according to parameter $softmax$

$$sIP(w) = e^{IP(w)S(w)}$$

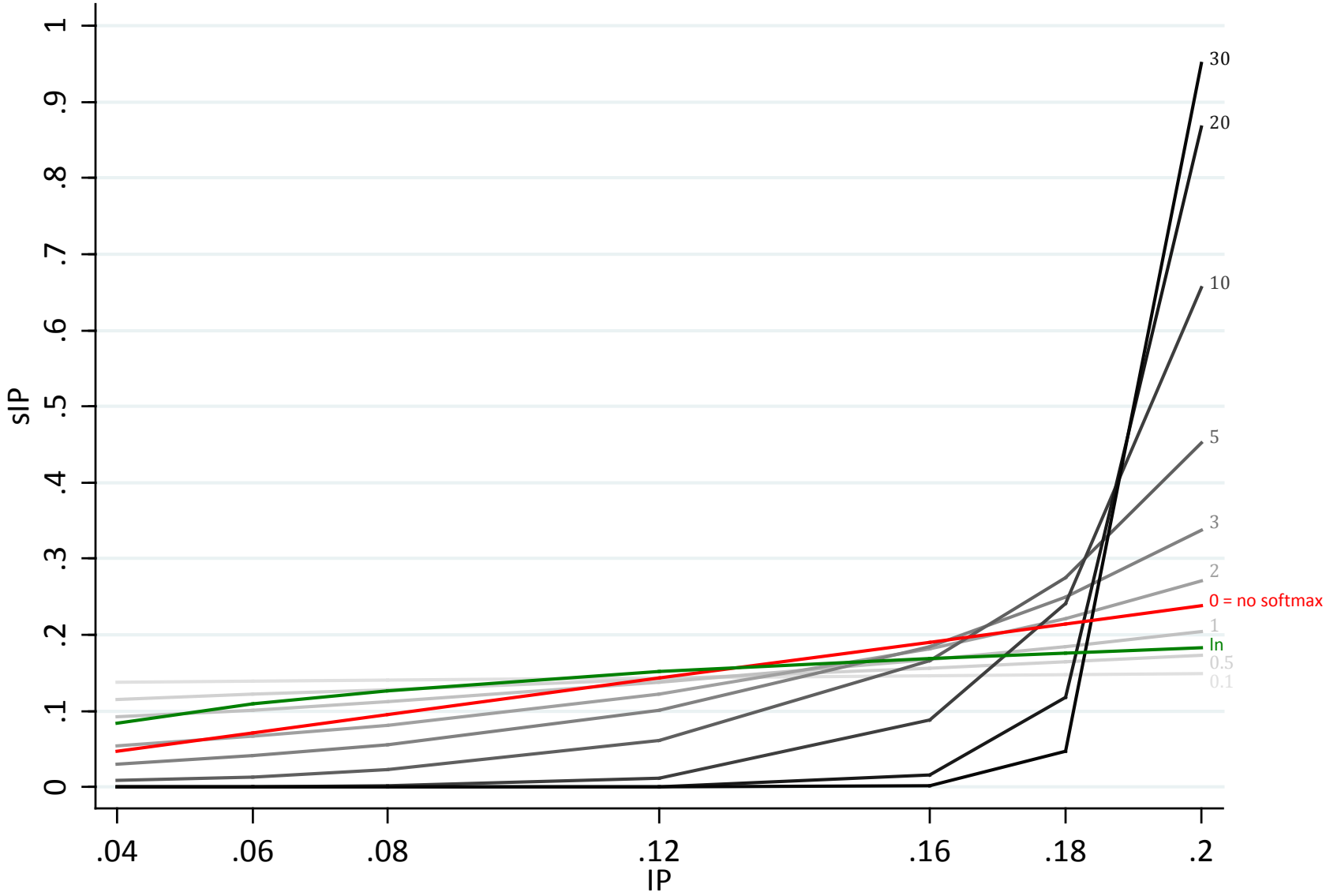
$$S(w) = \frac{sm\big(st(w)\big)}{\max\big(IP(v)|v \in S_{st(w)}\big)}$$

$IP(w)$ resolves to the absolute identification potential of word $w$
$S_t$ is the set of all words in the search term belonging to search type $t$
$st(w)$ returns the search type of word $w$
$sm(t)$ returns the $softmax$ parameter for search type $t$:

- $0$ : no softmax
- $0 < softmax \leq 1$ : smoothing
- $1 < softmax \leq 3$ : weak smoothing/accentuating
- $3 < softmax \leq 30$ : accentuating

Typical values: 0.1, 0.5, 3, 5, 9, 12, 20

# Softmax effect



# Paul Jaccard, 1868-1944, professor of botany and plant physiology, ETH Zurich

Searched Flower

Found Flower

$$J(S,F) = \frac{|\{S \cap F\}|}{|\{S \cup F\}|}$$

4 / 7 + 3 feedback

# Feedback $f$ as a slide control

$$Jaccard(w) = \frac{\sum_{v \in S} \begin{cases} IP(v) & | \; st(w) = st(v) \\ 0 & | \; st(w) \neq st(v) \end{cases}}{\sum_{v \in S \cup F} \begin{cases} IP(v) & | \; st(w) = st(v) \\ 0 & | \; st(w) \neq st(v) \end{cases}}$$

$$rIP(w, f) = rIP(w)\big((1 - f) + Jaccard(w)f\big)$$

$S$ is the search term (set of searched words)
$F$ is the candidate term (set of found words)
$f$ is the feedback factor: $f \in [0,1]$
$w$ is a word of the search term: $w \in S$
$st(w)$ returns the search type of word $w$
$IP(w)$ resolves to the absolute identification potential of word $w$
$Jaccard(w)$ returns the Jaccard coefficient of word $w$
$rIP(w, f)$ returns the relative identification potential for word $w$ with feedback $f$

# Feedback

Usually, feedback is applied subsequently on candidates retrieved by the basic heuristic

- General setting
- Ranking of candidates ($f \leq 0.1$)
  - Sorts the candidates by relevancy of additional noise without compromising the threshold
- Enforcing commutativity and transitivity ($f = 1$)
  - Transformation into an weighted Jaccard index
  - In combination with offset smoothing ($off(t) \leq -maxocc(t)$) all words of type $t$ get a weight of 1 → pure Jaccard index
- Softmax smoothing/accentuating will be ignored for surplus words as its inclusion would require a recalculation of all rIP not only the surplus ones.

# Cutoff and activation

## A high number of candidates for a search term is an indicator for redundancy

- General setting
- If the number of candidates for a given search term exceeds the cutoff…
  - sort candidates in descending order by identity ($\sum rIP$)
  - register the identity at the cutoff$^{th}$ position
  - remove all candidates with a lower identity
- Cutoff is an arbitrary threshold based on experience with the data that can still lead to huge candidate lists if there is no variance among them
- Additionally, temporary feedback can be applied to induce variance.
  If the number of candidates exceeds the activation limit…
  - apply feedback onto the candidates
  - apply cutoff procedure
  - remove feedback to prevent inconsistent identities in the result table

# Historical data and variants

- Export option
- Sources of variants
  - Some data sources contain by definition historical information, e.g. patent data
  - Databases with dedicated firm focus may provide historical information of name or address changes
  - Sometimes it is possible to create additional artificial variants for the same firm by…
    - adding a department name
    - separating expression in brackets assuming abbreviations of the firm name
    - suppressing numbers, joining words separated by hyphens (bio-tec → biotec), …
- Differentiation between unique key and group key separately for search and base table
  - Unique key unambiguously identifies a record (can be a simple record number)
  - Group key identifies an entity, e.g. a firm represented by multiple unique keys of variants
- Export of the best matches between entities of the base and search table among the retrieved candidates
  → suppressing redundancy without loosing information

# Incremental search strategies

## Develop search strategies based on multiple search runs

- Search types with a weight of zero will be ignored
- Concentrate the weights on search types linked to context defining search fields, like names, distributing the rest on the auxiliary types, like addresses (i.e. 70% name, 30% address)
- First search steps should be without destructive preparer, playing with different weight distributions for the auxiliary types

… followed by dedicated runs to hunt misspellings, setting the weights for corresponding non-destructive types to zero, e.g. 0% for basic name search type, 70% for name with gram-3 preparer

- Destructive preparer should only be applied to main fields not to auxiliary fields to save resources (time)
- Always have two runs to hunt misspellings: one with log smoothing for the destructive types and one without to account for the disruptive nature of these artefacts
- Think about which weight combinations of search types will make it over the threshold

# Disambiguation: search table = base table

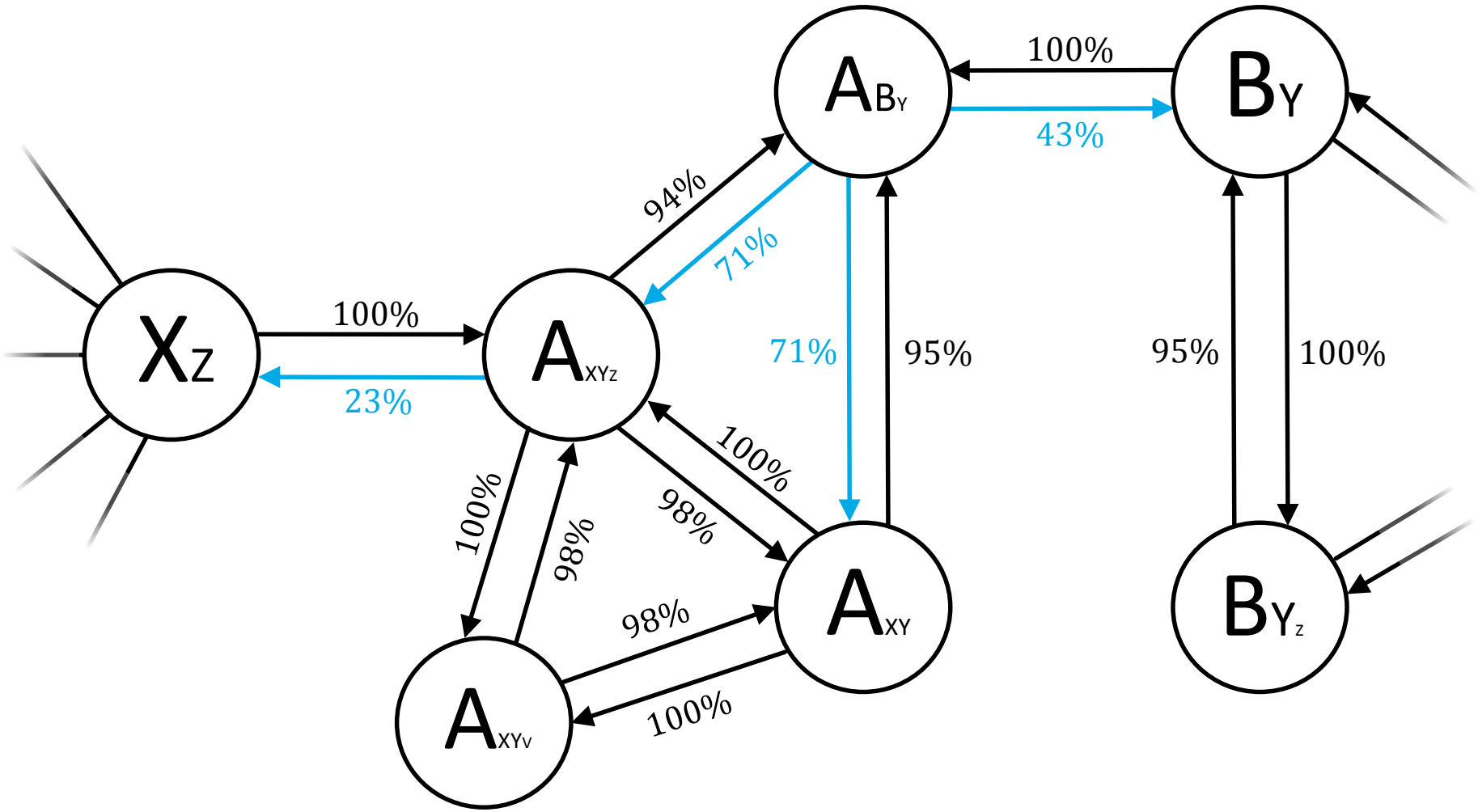
| searched | found | identity | equal | name | zip | city | street | country |
|---|---|---|---|---|---|---|---|---|
| 203346 | | | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | |
| 203346 | 203346 | 100.00 | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| | | | | | | | | |
| 71132 | | | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | |
| 71132 | 71132 | 100.00 | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | DE |
| 71132 | 66486 | 95.00 | | HOECHST SCHERING AGR EVO GMBH | 13347 | BERLIN | GERICHTSTRASSE 27 | DE |
| 71132 | 73565 | 90.00 | | HOECHST SCHERING AGR EVO GMBH | 13509 | BERLIN | MIRAUSTRASSE 54 | DE |
| | | | | | | | | |
| 323602 | | | | SCHERING AG | 13342 | BERLIN | MUELLERSTRASSE 170 178 | |
| 323602 | 323602 | 100.00 | | SCHERING AG | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 389129 | 99.57 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 199658 | 99.29 | | SCHERING AKTIENGESELLSCHAFT | 13342 | BERLIN | PATENTE MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| 323602 | 203346 | 99.29 | | SCHERING AKTIENGESELLSCHAFT PATENTE | 13342 | BERLIN | MUELLERSTRASSE 178 POSTFACH 65 03 11 | DE |
| 323602 | 402998 | 95.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 323602 | 180193 | 94.73 | | SCHERING AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 303857 | 94.73 | | SCHERING AG | 13353 | BERLIN WEDDING | MUELLERSTRASSE 178 | DE |
| 323602 | 397563 | 94.73 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 264 | 94.57 | | SCHERING AKTIENGESELLSCHAFT | 13342 | BERLIN | | DE |
| 323602 | 71132 | 94.57 | | HOECHST SCHERING AGR EVO GMBH | 13342 | BERLIN | GERICHTSTRASSE 27 | DE |
| 323602 | 171208 | 94.29 | | SCHERING AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 323602 | 435123 | 94.29 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICIENSING MUELLERSTRASSE 178 | DE |
| | | | | | | | | |
| 402998 | | | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | |
| 402998 | 402998 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 402998 | 397563 | 99.73 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 402998 | 435123 | 99.40 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICIENSING MUELLERSTRASSE 178 | DE |
| 402998 | 389129 | 94.68 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| | | | | | | | | |
| 518249 | | | | BAYER PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | |
| 518249 | 518249 | 100.00 | | BAYER PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 397563 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 402998 | 100.00 | | BAYER SCHERING PHARMA AG | 13353 | BERLIN | MUELLERSTRASSE 170 178 | DE |
| 518249 | 435123 | 98.76 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | PATENTS LICIENSING MUELLERSTRASSE 178 | DE |
| 518249 | 441578 | 98.76 | | BAYER PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLERSTRASSE 178 | DE |
| 518249 | 543908 | 94.52 | | BAYER PHARMA AKTIENGESELLSCHAFT | 13353 | BERLIN | MUELLER STRASSE 178 | DE |
| 518249 | 389129 | 93.76 | | BAYER SCHERING PHARMA AKTIENGESELLSCHAFT | 13342 | BERLIN | MUELLERSTRASSE 170 178 | DE |

# Disambiguation

- Export option
- Self referential search : base table = search table
- High threshold, e.g. 90%
- Not commutative, e.g. A→B above but A←B below threshold
  - Transformation to undirected edges A↔B with following quality attributes:
    - Minimum and maximum identity: min, max
      min is 0 if reverse linkage does not exists (below threshold)
    - Minimum score and percentile thereof: s, p
      missing reverse linkage is not considered
- Mirroring enforces identities and scores for non-existent reverse links (A←B)
  - Mirroring does not cause additional transitivity because no new edges are created
  - Reveals what is below the threshold to complete the min attribute of an edge which would otherwise be zero
  - Typically, the identity of mirrored links are based on smoothing to distribute the identification over all words:
    Even if the defining word is missing, there can be still a high similarity pertaining the other words. The defining word may be an artifact.
  - Optional

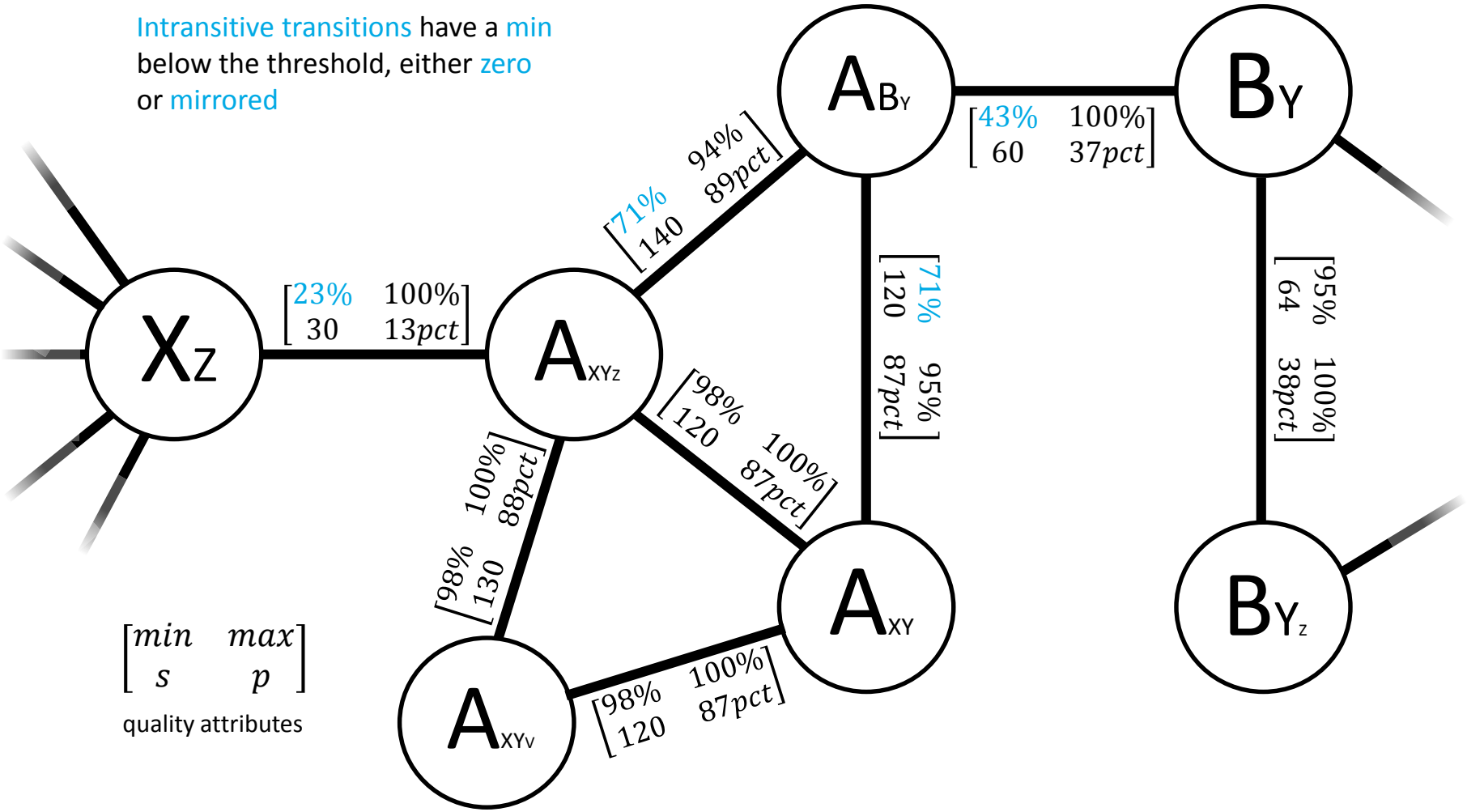
# Graph before transformation

Blue lines are mirrored (reverse identity below threshold of 90%)

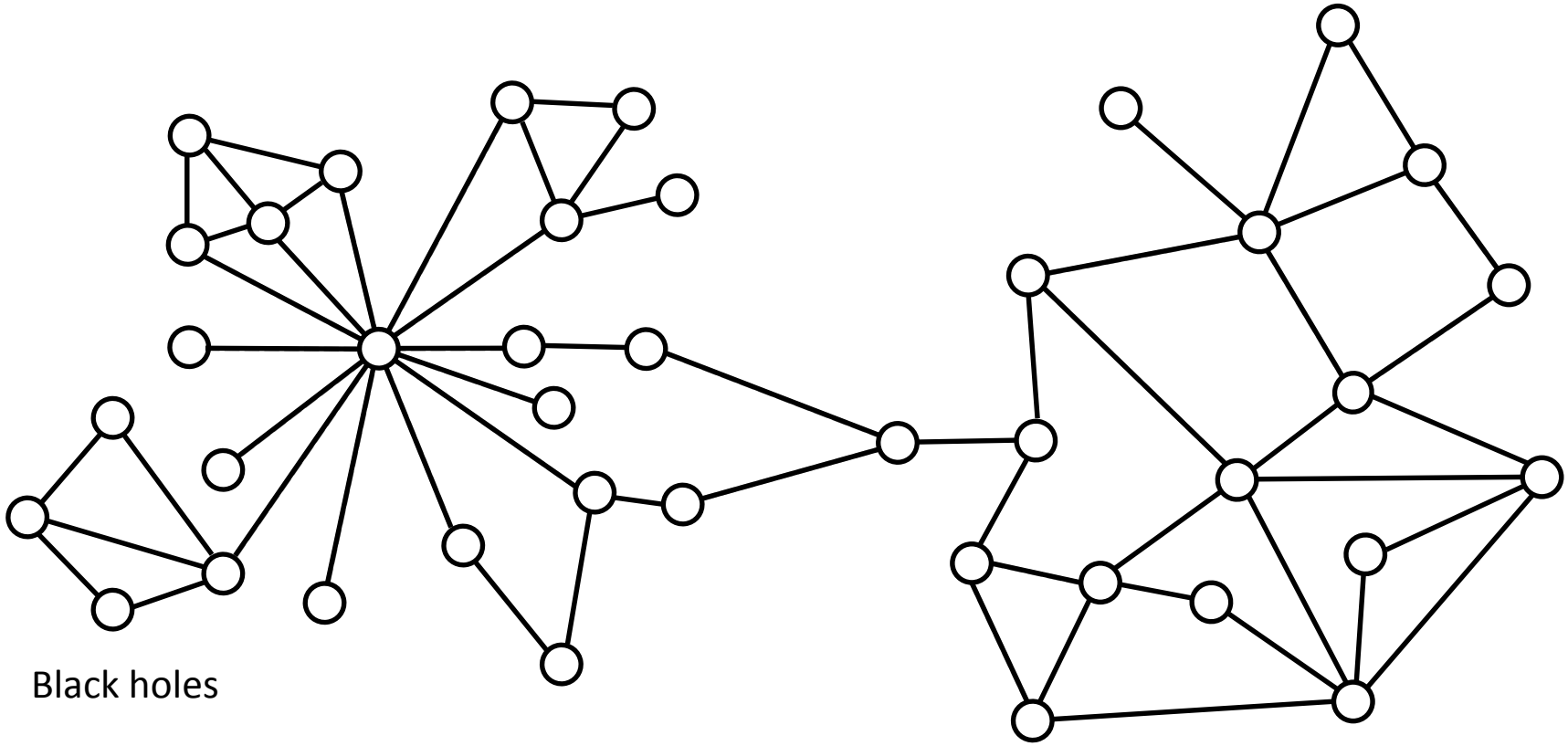


# Traversal

## Find all reachable nodes from a given starting node to define a cluster



# Clusters of unusual size



Black holes

- Data artefacts lead to suspiciously large number of connections

→ Cutoff and activation reduce impact

→ Artefact threshold as part of cascaded traversal curtails all connections

Thickets

- Only detectable during traversal

→ Thinning out the thicket by cutting weak connections: cascaded traversal

# Cascaded traversal

- Define a set of rules with increasingly restrictive conditions for the validity of a connection. Any rule has to be more restrictive than the previous rule.
- Attach a cluster threshold to every rule, e.g.:
  $unrestricted, \min \geq 90 \ @ \ 5, \min \geq 95 \ @ \ 11$
  $unrestricted, \mathrm{p} \geq 75 \text{ and } \min \geq 60 \text{ or } \min \geq 85 \ @ \ 6, \min \geq 90 \ @ \ 11$
- The rules will be exclusively activated in order of definition. The active rule will be replaced if the cluster threshold of the following rule is reached.
- Every time a new rule is activated, the traversal of the network starts again at the given start node with the new rule in place.
- A valid start node is any node that does not already belongs to a cluster created by another start node.
- Any rule creates a new virtual network that is a thinned out version of the network defined by the previous rule. As the propagation of this thinning out process is independent from the start node, there is no overlapping of the resulting clusters.
- Artefact threshold is defined by a single number before the first rule of a cascade, e.g.: $200, \min \geq 90 \ @ \ 5, \ldots$

# Edge types

- Weak edges
  - Given a high threshold, the maximum of an edge carries no additional information besides the link itself. A missing or low minimum is an indicator for a weak edge.
- Strong edges
  - Given a high threshold, a minimum of above the threshold constitutes a strong edge based on an almost commutative relation as part of a transitive cluster, e.g. only irrelevant variations like different legal forms between firm A and B
  - A mirrored minimum below but close to the threshold in conjunction with a relatively high score percentile represents a strong edge, e.g. well identified firm → joint venture with global corporation
  - Even without mirroring, a high score percentile of an edge, which is based on the minimum of both involved scores, is a good indicator for at least some mutual important components, e.g. well identified firm → firm + exotic subsidiary information

→ Given a high threshold, most rules are based on the minimum min and the score percentile p

# Defining a cascade

## Assessment of the variation of an entity in the data

- Unrestricted
  - In the unrestricted zone everything can happen: unmonitored intransitive transitions
  - How many variants of an entity are still plausible?
    Answer: 4 → $unrestricted, \min \geq 90 \ @ \ 5$
  - First rule: $\min \geq threshold \ @ \ plausible + 1$
  - Subsequent rules allow for higher activation limits according to the increased quality of the remaining links → transitivity is already enforced
- Restricting the unrestricted
  - Imposing rules on the unrestricted zone below the threshold: controlled intransitive transitions
  - With mirroring: rules can incorporate min (below threshold) and score percentile, e.g.:
    $\mathrm{p} \geq 80 \text{ or } \min \geq 75 \ @ \ 0, \min \geq 90 \ @ \ 5$
    $\mathrm{p} \geq 70 \text{ and } \min \geq 60 \text{ or } \min \geq 80 \ @ \ 0, \min \geq 90 \ @ \ 5$
  - Without mirroring: rules are only based on score percentile, e.g.:
    $\mathrm{p} \geq 90 \ @ \ 0, \min \geq 90 \ @ \ 5$
- All rules include an implicit: … and $\max \geq threshold$ …

# Nested cascaded traversal

## Heterogeneous structure in the data leads to conflicting interests

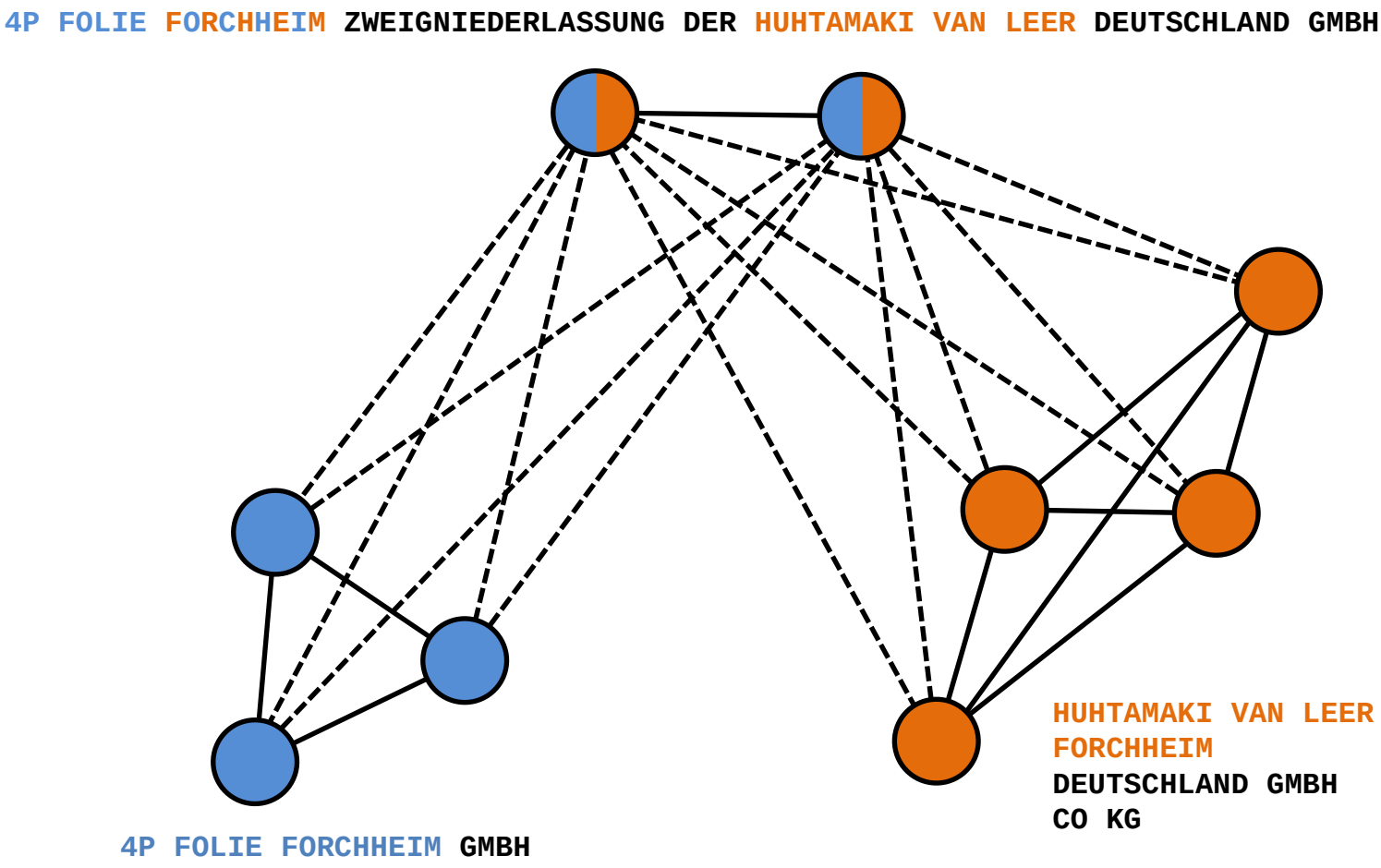
- Short entries are less prone to variation than their long and more complex counterparts:
  Secretary of State for Trade and Industry in Her Britannic Majesty's Gov. of the U.K. of Great Britain
  *Mayer GmbH*
  - Long entries would require larger cascade thresholds due to higher variation
  - … exceeding the optimal size for *shorter* entries

→ Two nested cascade runs

1. Run with a transitivity enforcing rules at cluster threshold zero.
   Pre-clustering of very similar, almost transitive entries harmonizes variation.
   The rule defines the new maximum of the involved quality attributes.
2. Run with cluster thresholds attuned to a more homogenous network.
   How many major changes are still plausible?
   How many intransitive transitions am I willing to accept?
   - Examples:
     $\min \geq 97 \ @ \ 0$; $\min \geq 90 \ @ \ 4, \min \geq 95 \ @ \ 7$
     $p \geq 75$ and $\min \geq 80$ or $\min \geq 97 \ @ \ 0, \min \geq 97 \ @ \ 201$; $\min \geq 90 \ @ \ 4, \min \geq 95 \ @ \ 7$
     $\min \geq 90 \ @ \ 0$; $p \geq 75$ and $\min \geq 80 \ @ \ 4, \min \geq 100 \ @ \ 8$

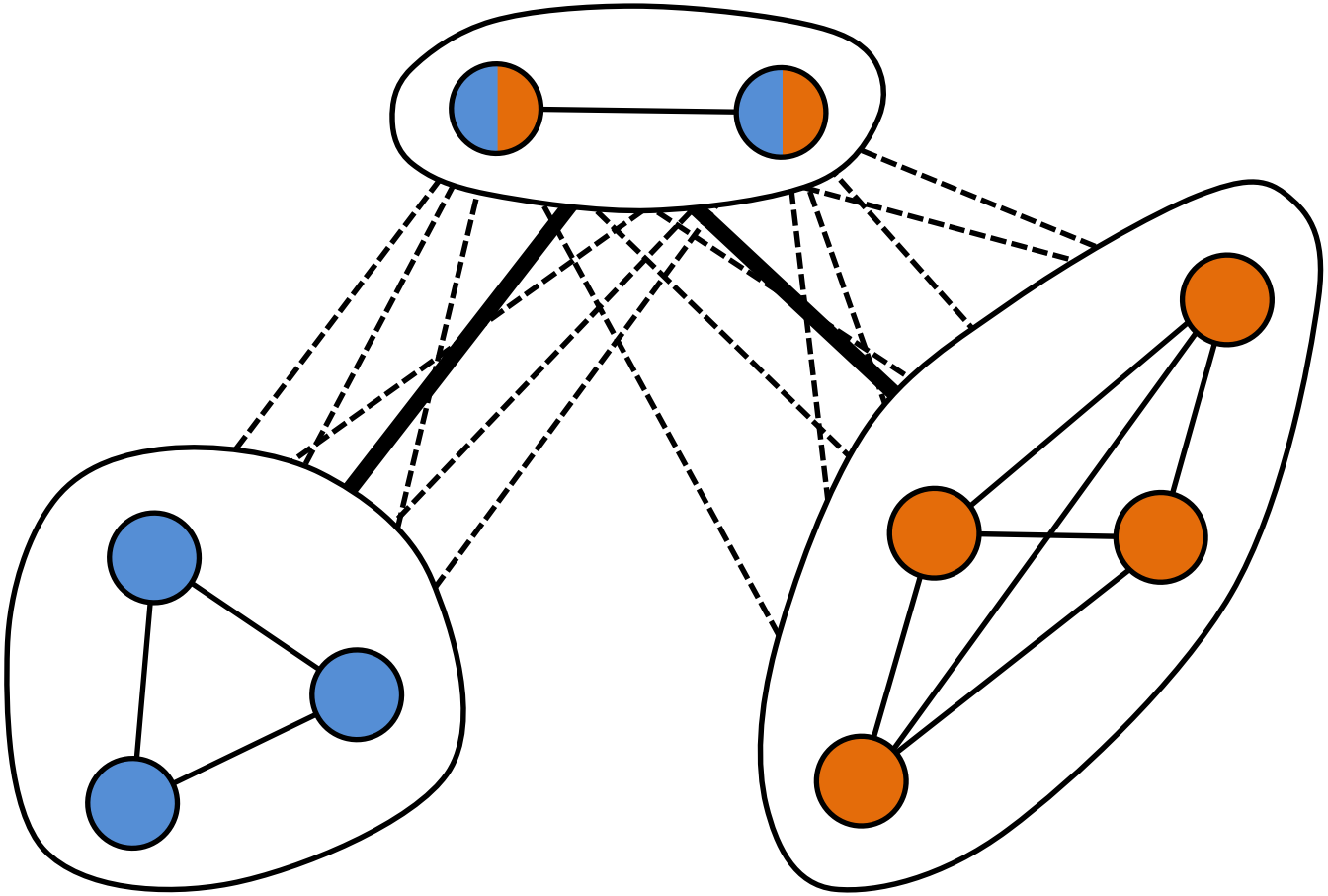
- Doesn‘t apply to data with less contextual noise, i.e. person names

# Nested cascade (before pre-clustering)



# Nested cascade (after pre-clustering)

## Aggregation: maximizing quality attributes of connections



# A short cascade guide

- Try several cascades to improve the ruleset. Usually, there is no benefit in overly complex rules. Keep different results for different purposes: coarse (high cluster thresholds) if additional restriction minimize false positives and fine (low thresholds) otherwise.
- High scores are an indicator for search terms with a high identification potential and deserve a more lenient handling, e.g. first ruleset of a nested cascade with fallback: $p \geq 85$ and $min \geq 60$ or $min \geq 90$ @ $0$, $min \geq 90@101$;…
- Nested cascades: try first ruleset separately if score-based properties are involved → better assessment of follow-up cascade
- A normal result output (without clustering) gives a good insight into the specific data issues and the score distribution
- Mirrored searches return a less skewed distribution if log smoothing is applied (just like in econometrics) → improves the interpretability of the min property
- Clusters do not have to and can not be perfect. Many applications, like identification of self-citations in patents, enforce additional restrictions minimizing the impact of inflated clusters. Just avoid ending up with one super-cluster.

https://github.com/ThorstenDoherr/searchengine

Use at your own risk.
Thank you for your attention.