Network Metrics for Assessing the Quality of Entity Links Between Multiple Datasets

Al Idrissou[1,2], Frank van Harmelen[1], and Peter van den Besselaar[2]

[1] Department of Computer Science, Vrije Universiteit Amsterdam, NL
[2] Department of Organization Sciences, Vrije Universiteit Amsterdam
{o.a.k.idrissou,frank.van.harmelen,p.a.a.vanden.besselaar}@vu.nl

Abstract. Linking entities between datasets is a crucial step in data-integration in general, and in the use of multiple datasets on the semantic web in particular. A rich literature exists on different approaches to the entity linking problem, and a fair amount of tools is available for practical use. However, much less work has been done on how to assess the quality of such entity links once they have been generated by any of these tools. Evaluation methods for link quality are typically limited to either comparison with a ground truth (which is often not at one's disposal), manual work (which is cumbersome and prone to error), or crowd sourcing (which is not always feasible, especially if background information is required). Furthermore, the problem of link evaluation is greatly exacerbated for links between more than two datasets, because the number of possible links grows rapidly with the number of datasets.
In this paper we propose a method to estimate the quality of such entity links between multiple datasets. We exploit the fact that the links between entities from multiple datasets form a network, and we show how simple metrics on this network of entity-links can reliably predict the quality of these links. We verify our results in a large experimental study using six datasets from the domain of science and innovation studies.

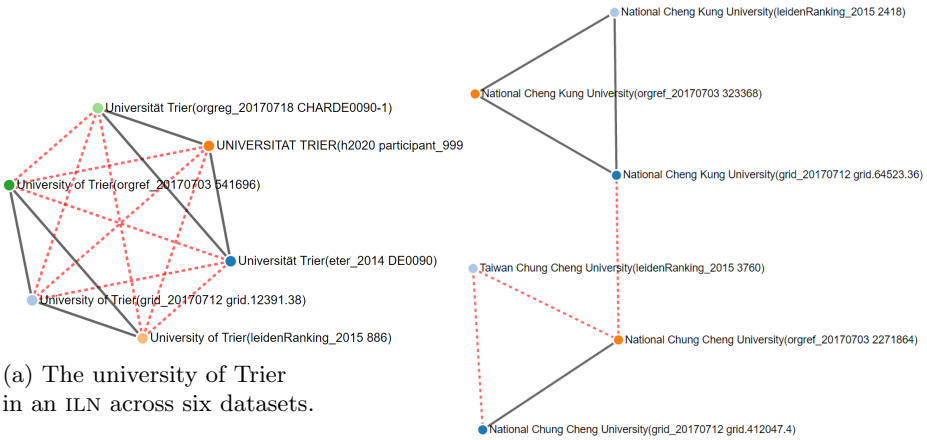
Keywords: entity linking, data integration, network metrics

1 Introduction

Linking entities between datasets is a crucial step in data-integration in general, and in the use of multiple datasets on the semantic web in particular. However, the challenge in the semantic web community is not only the availability of methods for linking resources (there exist a fair amount of matching tools: AGDISTIS [3], Linkage Query Writer [1, 2], SILK [4], etc.), but it is also how to validate the links produced by theses tools. At the moment, only three validation options are available: (1) *ground truth*, but a validated base line is not often at one's disposal; (2) *manual work*, which is a cumbersome task prone to error; (3) *crowd sourcing*, which is not always feasible especially if (specialist) background information is required. Furthermore, the problem of link evaluation is greatly exacerbated for links between more than two datasets, because the number of possible links grows rapidly with the number of datasets. This suggests that it is of importance to investigate *"how to efficiently help users to more accurately*

evaluate discovered links in a shorter time frame?". Any answer to this question should generalise beyond the setting of just two datasets, and be applicable instead to the general setting of links between multiple datasets. In such a multi-dataset scenario, linked resources cluster in small groups that we denote *Identity Link Networks (ILNs)*. The goal of this paper is to provide users with a way to estimate the quality of such identity link networks, and consequently validate a set of discovered links at once. To do so, *we hypothesize that the structure of such an identity link network correlates with its quality.*

We test our hypothesis through two experiments where we show that the proposed metrics indeed helps to estimate the quality of an identity-network with a level of certainty that can be derived from the strength of the weakest link in the identity-network.



(a) The university of Trier in an ILN across six datasets.

(b) Potentially wrong representation of the National Chung Cheng University

Fig. 1: Two examples of Identity Link Networks (ILN); dotted lines indicate links with a low confidence.

2 Identity Link Network

We assume the well known setting of a real-world entity that has one or more digital representations in multiple datasets. The task of entity linking is to discover which entity (or entities) in each dataset denotes the same real world entity. An Identity Link Network (ILN) is a network of links between entities from a number of datasets that are found by one or more linking algorithms to represent the same real world entity.

Figure 1 shows two examples of such ILNs that have been generated by an entity linking algorithm between entities from 6 datasets. Fig. 1a shows the

ILN for the real world entity "University of Trier", fig 1b shows the same for the National Chung Cheng University. *In this paper, we hypothesise that the structure of these ILNs is a reliable indicator for the correctness of links in the network.*

3 Network Properties & the Quality of a Link-Network.

Figure 2 illustrates a set of seven simple network topologies using the same number of nodes. Depending on the problem at hand, different network topologies are preferred, since different network topologies can have very different features and properties to them. Part of our proposed metrics is based on the intuition that multiple links provide corroborating evidence for each other, suggesting that in the case of an ILN, the ideal topology is a ***fully connected*** network. It illustrates a *total agreement* between all resources (not the case for any other topology), and it *does not require any intermediate* resource to validate the existence of an identity-link between any other two resources (again, not the case for any other topology). We will capture these and similar intuitions in three different quantitative metrics over ILNs: *Bridge*, *Diameter* and *Closure*.

In the next paragraphs of this section, we first define and explain the rationale behind each metric, then normalise each measure to output values between 0 (*no impact*) and 1 (*maximum impact*), and finally average the sum of all metrics as the overall quality of the network which we will use for estimating the quality of the ILN.

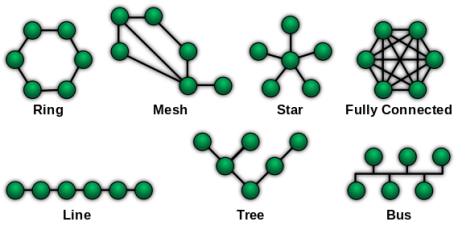


Fig. 2: Example of network topologies.
Source: https://en.wikipedia.org/wiki/Network_topology

Bridge. A bridge[3] in a graph is an edge whose removal causes the number of connected components of the graph to increase. Equivalently, a bridge is an edge that does not belong to any cycle. The intuition for this measure is that the presence of a bridge in an ILN suggests a potentially problematic link which is not corroborated by any other links.

As a graph with n nodes can contain at most $max_b = n - 1$ bridges (e.g. in a ***Line*** network structure), the bridge value is normalised as $n_b = \frac{B}{max_b}$, where B is the number of bridges. An ideal link network would have no bridge ($n_b = 0$). As n_b is sensitive to the total number of nodes in the graph (it decreases for large graphs, even when the number of bridges is the same), we "soften" the value of n_b with a sigmoid function: $n'_b = max(n_b, sigmoid(B))$, where the sigmoid function $sigmoid(x) = \frac{x}{|x|+1.6}$ helps stabilising the impact of the size of the graph by providing a minimal value for n'_b. The value 1.6 is a hyper-parameter that has been determined empirically.

[3] also known as an isthmus or a cut-edge, https://en.wikipedia.org/wiki/Glossary_of_graph_theory_terms

Diameter. The diameter D of a graph with n nodes is the maximum number of edges (distance) in a shortest path between any pair of vertices (i.e. the longest shortest path). In an ideal scenario, if three resources A, B and C are representations of the same real world object, there would be no need for an intermediate resource for confirming the identity of any of the resource in the network. For example, if determining that A and C are the same is only possible by using B, then there is a need for checking whether the newly acquired knowledge (A *:sameAs* C) is indeed accurate or whether there is a missing shared property between A and C that causes them not to be linked. In a fully connected graph of n nodes, the diameter $D = 1$. The longest diameter is observed in a ***Line*** network structure, with $max_d = n - 1$ for a line network of n nodes. To scale to the [0,1] interval, the diameter is normalised as $n_d = \frac{d-1}{max_d - 1}$. Like the bridge, because the diameter is also sensitive to the number of nodes, the normalised diameter is calculated as $n'_d = max(n_d, sigmoid(D - 1))$.

Closure. In a connected graph of n nodes, the closure is the ratio of the number of observed arcs A over the total number of possible arcs $max_a = \frac{1}{2}n(n-1)$. In a complete graph, this ratio has value 1. Hence, to evaluate how far the observed graph is from the ideal (complete) one, we normalise the closure as $n_c = 1 - \frac{A}{max_a}$. The minimum number of connections is $min_a = (n - 1)$, as observed in ***line*** and ***star*** network structures.

Estimated Quality. All of these metrics capture the same intuition: the more an ILN resembles a fully connected graph, the higher the quality of the links in the ILN. Of course, these three metrics are not independent: $n_c = 0$ implies $n'_d = 0$ and $n'_b = 0$. However, using only n_c would be too uninformative (e.g. the converse of the implication does not hold), and each of n_c, n'_d and n'_b capture different (though related) amounts of redundancy in the ILN. Consequently, to compute an overall estimated quality e_Q of an identity link network, we simply average the three separate metrics, and invert them so that the value 1 indicates the highest quality: $\mathbf{e_Q} = 1 - \frac{n'_b + n'_d + n_c}{3}$.

Discrete Warning Intervals. The e_Q metric scores all ILNs on a continuous value in the [0,1] interval. In order to automatically discriminate potentially good networks from bad ones, we divide this interval into three segments: ILNs with values $0.9 \leq e_Q \leq 1$ will be rated as good, with values $0.75 < e_Q < 0.9$ as undecided, and with values $0 \leq e_Q \leq 0.75$ as bad. These interval boundaries have been empirically determined, and can of course be adjusted depending on the use case,

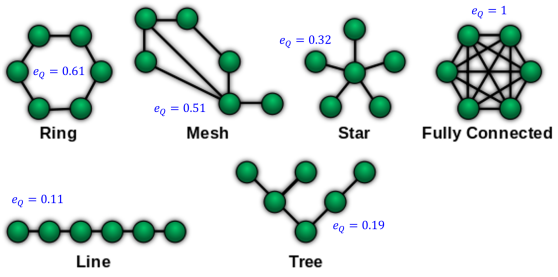


Fig. 3: e_Q values on example networks

where some use-cases may be more tolerant of low quality links than others. The specific values of these boundaries does not affect the essence of our hypothesis.

Hypothesis. We can now state our hypothesis more formally:
the e_Q intervals defined above are predictive of the quality of the links in an entity link network between multiple datasets.

Example. By way of illustration, figure 3 shows the value of our metric e_Q for the six networks from Figure 2, showing that our metric e_Q does indeed capture redundancy in a network.

In the following sections, we will test this hypothesis against human evaluation on hundreds of ILNs containing thousands of links between 6 datasets. Section 4 describes the datasets involved in our experiment, sections 5 and 6 describe our two experiments, and section 7 concludes.

Link-Network Quality Estimation				
ILN	Bridge $n_b' = B/max_b$ $n_b = max(n_b', sig(B))$	Diameter $n_d' = (D-1)/(max_d - 1)$ $n_d = max(n_d', sig(D-1))$	Closure $n_c = 1 - A/max_a$	Est. Quality $e_Q = 1 - impact$
Ring	$b = 0 \quad n_b = 0.00$	$d = 3 \quad n_d = 0.56$	$c = 0.40 \;\; n_c = 0.60$	$\mathbf{e_Q = 0.61}$
Mesh	$b = 1 \quad n_b = 0.38$	$d = 3 \quad n_d = 0.56$	$c = 0.47 \;\; n_c = 0.53$	$\mathbf{e_Q = 0.51}$
Star	$b = 5 \quad n_b = 1.00$	$d = 2 \quad n_d = 0.38$	$c = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q = 0.32}$
Full Mesh	$b = 0 \quad n_b = 0.00$	$d = 3 \quad n_d = 0.00$	$c = 1.00 \;\; n_c = 0.00$	$\mathbf{e_Q = 1.00}$
Line	$b = 5 \quad n_b = 1.00$	$d = 1 \quad n_d = 1.00$	$c = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q = 0.11}$
Tree	$b = 5 \quad n_b = 1.00$	$d = 4 \quad n_d = 0.38$	$c = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q = 0.34}$

Table 1: Metrics for analysing and computing the quality of a network of links.

4 Datasets

Our datasets are taken from the domain of social science, more specifically from the field of Science and Innovation Studies. Entities of interest to this domain of study are (among others) universities and other research-related organisations, such as R&D companies and funding agencies. Our 6 datasets are widely used in the field, and describe organisations and their properties such as name, location, type, size and other features[4].

Grid[5] describes 80248 organisations across 221 countries using 12308 relationships. Only 17 countries (United States, United Kingdom, Japan, Germany, France, Canada, Czechia, China, India, Norway, Italy, Spain, Brazil, Russia, Switzerland, Sweden and Australia) within Grid host a thousand or more organisations. This accounts for 77% of the total. All organisations within Grid are

[4] The information provided here about the datasets was collected in January 2018. The datasets themselves are of earlier dates: Grid: 2017.07.12; Orgref: 2017.07.03; OpenAire: 2018.08.16; OrgReg: 2017.07.18; Eter: 2014; Leiden Ranking 2015: 2017.6.16; and Cordis-H2020: 2016.12.22.

[5] `https://www.grid.ac`

assigned an address, while 96% of them have an organisation type (company, education, healthcare, non-profit, facility, government, archive, and 'other'), and only 78% have geographic coordinates.

OrgRef[6] collates existing data about the most important worldwide academic and research organisations (31000) from two main sources: Wikipedia and ISNI. The following types of institutions are distinguished: universities, colleges, schools, hospitals, government agencies and companies involved in research.

The Leiden Ranking dataset[7] offers scientific performance indicators of more than 900 major universities. These universities are only included when they are above the threshold of 1000 fractionally counted Web of Science indexed core publications. This explains its coverage across only 54 worldwide countries.

Eter[8] is a database on European Higher Education Institutions that not only includes research universities, but also colleges and a large number of specialized schools. The dataset covered 35 countries in 2015.

OrgReg[9] is based on ETER but adds to the about 2700 HE institutions some 500 public research organizations and university hospitals. Collected between 2000 and 2016, its organisations are distributed across 28 European countries, 2 EEA-EFTA countries (Iceland, Liechtenstein, Norway and Switzerland), as well as four candidate countries (FYRM, Montenegro, Serbia and Turkey)

The European Organisations' Projects H2020 database[10] documents the projects the participating organisations from Horizon 2020, the largest EU Research and Innovation programme with the focuses on excellent science, industrial leadership and societal challenges.

5 Experiment 1

We test our hypothesis on a real life case study that revolves around six datasets, with as goal to investigate the coverage of OrgReg (coverage analysis of datasets is a typical question asked by social scientists before including a dataset in their studies). This is done by comparing the entities in OrgReg to those in Grid, OrgRef, the CWTS Leiden Ranking, ETER, and H2020.

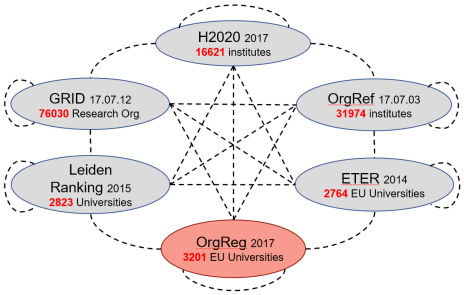


Fig. 4: Disambiguating OrgReg

[6] http://www.orgref.org
[7] http://www.leidenranking.com/
[8] https://www.eter-project.com/
[9] http://risis.eu/orgreg/
[10] http://www.gaeu.com/sv/item/horizon-2020

5.1 Description

Using the organisations' name, we create 21 sets of links between each pair of datasets (including linking a dataset to itself in order to detect duplicate entities in the dataset).

Organizations are linked across or within datasets using approximate string matching on their names with minimal similarity threshold 0.8. We then take the union of all 21 sets of links, resulting in a collection of ILN's of varying size (see figure Figure 5).

Now that we have constructed a large collection of multi-dataset ILNs, we will compute the e_Q value for all ILNs, and then check the predicted good/bad categories against two human experts.

5.2 Results of first evaluation

Ideally, we would find only ILNs of size 6 when each Orgreg entity is linked with one and only one entity in each of the 5 other datasets. With less than 100% coverage of Orgreg, we expect to find ILNs of size < 6. Fig. Figure 5 shows that we also find a substantial number of ILNs of size > 6. This is due to (a) duplicates occurring in a single dataset, resulting in links in the ILN between two items from the same dataset, and (2) an imperfect matching algorithm (in our case approximate name matching), resulting in incorrect links in the ILN.

This shows that it is indeed important for a social scientist to judge the quality of the many thousands of ILNs, justifying the goal of this paper to provide the data user with a machine-calculated prediction of the quality.

Due to the high number of ILNs generated, we evaluate only the 822 ILNs of size 5 to 10, with the following frequencies: 391 (size 5), 224 (6), 96 (7), 66 (8), 45 (9) and 24 (10). We predict a 'good' or 'bad' score based on the e_Q value for each of these 822 ILNs, and then compare these scores against those of a human expert, resulting in recall, precision and F1 scores. In red, Figure 5 displays the F_1 measure for each ILN size. Overall, our e_Q metric resulted in high F_1 values ($0.806 \leq F_1 \leq 0.933$). These F_1 values are derived from a detailed confusion matrix[11] for each ILN size. As an example, Table 2 shows the confusion matrix for ILNs of size 8 which score the lowest F_1 value (F_1=0.806, based on 0.84 precision and 0.77 recall).

We also pitched our e_Q metric against a Majority Class Classifier. Table 3 shows that our e_Q metric outperforms the Majority Class Classifier on both F_1 measure, Accuracy (ACC) and Negative Predicted Value (NPC) for ILNs of all sizes.

All of these findings show the very strong predictive power of our e_Q metric for the quality of ILNs when compared to human judgement.

5.3 Results of second evaluation

For a further evaluation by a Dutch domain expert, we selected **148** ILNs (ranging from size 3 to 10 as depicted in Table 4) in which at least one resource is located

[11] https://en.wikipedia.org/wiki/Confusion_matrix

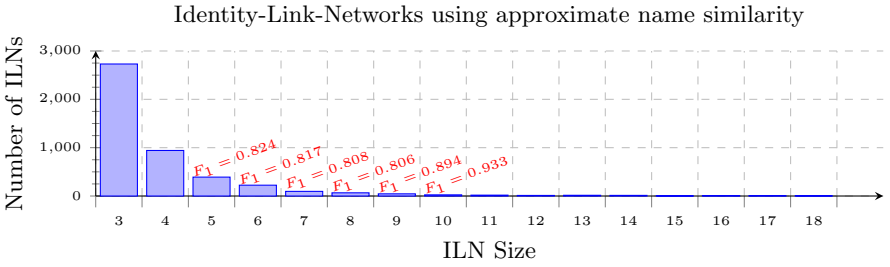


Fig. 5: Clusters extracted from the lens.

		66 GROUND TRUTHS			
		GT. Pos. 35	GT. neg. 31		
PREDICT	POSITIVE 32	True Pos. 27	False Pos. 5	Precision 0.844	False Discovery Rate 0.156
	NEGATIVE 34	False Neg. 8	True Neg. 26	F. Omission Rate 0.235	Neg. Predictive Value 0.765
		Recall 0.771	Fall-out 0.161	Positive L. Ratio 5.241	F1 score — Accuracy 0.806 — 0.803

Table 2: Confusion matrix for link-networks of size 8.

Majority Class Classifier vs Network Metrics $\frac{MajorityClassClassifier}{NetworkMetrics}$ GT = Ground Truth GT_P = Ground Truth Positive GT_N = Ground Truth Negative				
Size 5	GT_P=272 GT_N=119	F_1 : $\frac{0.821}{\mathbf{0.824}}$	ACC: $\frac{0.696}{\mathbf{0.747}}$	NPV: $\frac{-}{\mathbf{0.598}}$
Size 6	GT_P=139 GT_N=85	F_1 : $\frac{0.766}{\mathbf{0.817}}$	ACC: $\frac{0.621}{\mathbf{0.768}}$	NPV: $\frac{-}{\mathbf{0.709}}$
Size 7	GT_P=50 GT_N=56	F_1 : $\frac{0.685}{\mathbf{0.808}}$	ACC: $\frac{0.521}{\mathbf{0.792}}$	NPV: $\frac{-}{\mathbf{0.810}}$
Size 8	GT_P=35 GT_N=31	F_1 : $\frac{0.693}{\mathbf{0.806}}$	ACC: $\frac{0.530}{\mathbf{0.803}}$	NPV: $\frac{-}{\mathbf{0.765}}$
Size 9	GT_P=21 GT_N=24	F_1 : $\frac{-}{\mathbf{0.894}}$	ACC: $\frac{0.533}{0.889}$	NPV: $\frac{0.533}{\mathbf{1}}$
Size 10	GT_P=8 GT_N=16	F_1 : $\frac{-}{\mathbf{0.933}}$	ACC: $\frac{0.667}{\mathbf{0.958}}$	NPV: $\frac{0.667}{\mathbf{0.941}}$

Table 3: Comparing the network-metric result to the MCC baseline.

in the Netherlands. The Dutch expert deviated from the first evaluation in only 12 out of 148 cases (mainly where institutions shared a saint's name). These changes slightly affect the ground truth for each ILN size. The F_1 measures for these evaluations are even higher ($0.848 \leq F_1 \leq 1$) as compared to the previous experiment. However, the very imbalanced character of the ground truth makes it hard to always outperform the baseline as illustrated in Table 4. This second experiment confirms our finding in the first experiment that e_Q is a reliable predictor of ILN quality.

Majority Class Classifier vs Network Metrics $\frac{MajorityClassClassifier}{NetworkMetrics}$ GT = Ground Truth GT_P = Ground Truth Positive GT_N = Ground Truth Negative				
Size 3	GT_P=22 GT_N=2	F_1 : $\frac{\mathbf{0.933}}{0.931}$	ACC:$\frac{\mathbf{0.875}}{\mathbf{0.875}}$	NPV:$\frac{-}{\mathbf{0.5}}$
Size 4	GT_P=22 GT_N=2	F_1 : $\frac{\mathbf{0.884}}{0.878}$	ACC:$\frac{\mathbf{0.792}}{\mathbf{0.792}}$	NPV:$\frac{-}{\mathbf{0.5}}$
Size 5	GT_P=14 GT_N=1	F_1 : $\frac{\mathbf{0.966}}{0.929}$	ACC:$\frac{\mathbf{0.933}}{0.867}$	NPV:$\frac{-}{\mathbf{0}}$
Size 6	GT_P=13 GT_N=3	F_1 : $\frac{\mathbf{0.848}}{\mathbf{0.848}}$	ACC:$\frac{\mathbf{0.737}}{\mathbf{0.737}}$	NPV:$\frac{-}{-}$
Size 7	GT_P=11 GT_N=1	F_1 : $\frac{0.909}{\mathbf{1.0}}$	ACC:$\frac{0.833}{\mathbf{1.0}}$	NPV:$\frac{-}{\mathbf{1.0}}$
Size 8	GT_P=4 GT_N=0	F_1 : $\frac{\mathbf{1.0}}{\mathbf{1.0}}$	ACC:$\frac{\mathbf{1.0}}{\mathbf{1.0}}$	NPV:$\frac{-}{-}$
Size 9	GT_P=8 GT_N=1	F_1 : $\frac{0.941}{\mathbf{1.0}}$	ACC:$\frac{0.889}{\mathbf{1.0}}$	NPV:$\frac{-}{\mathbf{1.0}}$
Size 10	GT_P=1 GT_N=0	F_1 : $\frac{\mathbf{1.0}}{\mathbf{1.0}}$	ACC:$\frac{\mathbf{1.0}}{\mathbf{1.0}}$	NPV:$\frac{-}{-}$

Table 4: Comparing the network-metric result to the MCC baseline: Expert sample.

5.4 Analysis

Both of the evaluations of e_Q above (the evaluation by the authors of all ILNs of size 5-10, and the evaluation by a Dutch expert of all ILNs that involve a Dutch organisation) resulted in very high F_1 values (a weighted F_1 average of 0.826 in table Table 3 and 0.930 in Table 4). Furthermore, e_Q outperformed a majority-class classifier in Table 3 in the first experiment (not in the second because of the highly imbalanced distribution). All this supports our hypothesis that our e_Q measure is strongly predictive of the human judged quality of the links between the entities in an Identity Link Network.

6 Experiment 2

The previous experiment created links between entities using a rather weak matching heuristic, namely approximate string matching between entity names. This was an interesting setting because such weak linking strategies are a fact of daily life on the semantic web (and in data integration in general). In the next experiment, we will use e_Q to evaluate ILN's that has been constructed using a more sophisticated matching heuristic. We will see that also in this case, e_Q is strongly predictive of human judged link quality.

The stronger matching heuristic that we use in this second experiment combines organisation names with the geo-location of the organisation. The experiment is run over Eter, Grid and Orgreg as they are the only datasets at our disposal that document organisations with geo-coordinates.

6.1 Description

This subsection describes in three phases how the experiment is conducted.

Phase-1: Create links. The first phase links organizations across the three datasets (Eter-Grid, Eter-Orgreg and Grid-Orgreg) whenever they are located within an immediate vicinity of 50 meters, 500 meters and 2 kilometres. This creates nine sets of links (three for each vicinity).

Phase-2: Refine links. Each set of links is then refined by applying an approximate name comparison over the discovered linked resources with a threshold of 0.7. By now, we have sets of links without name comparison (geo-only) and with name comparison (geo+names), organised in three subgroups (50m, 500m and 2km) each.

Phase-3: Combine links. To generate the final ILNs, the sets of links within each subgroup are combined using the union operator. The goal of this is to compare within a specified distance, ILNs that where generated without name matching to those generated with name matching.

Experiment rationale. The reasons behind the increase of the proximity distance in this experiment is that by increasing the near-by distance we expect to increase the number of false positive links (noise) and we want to test if the e_Q metric will highlight potentially problematic ILNs.

6.2 Strict vs. Liberal Clustering

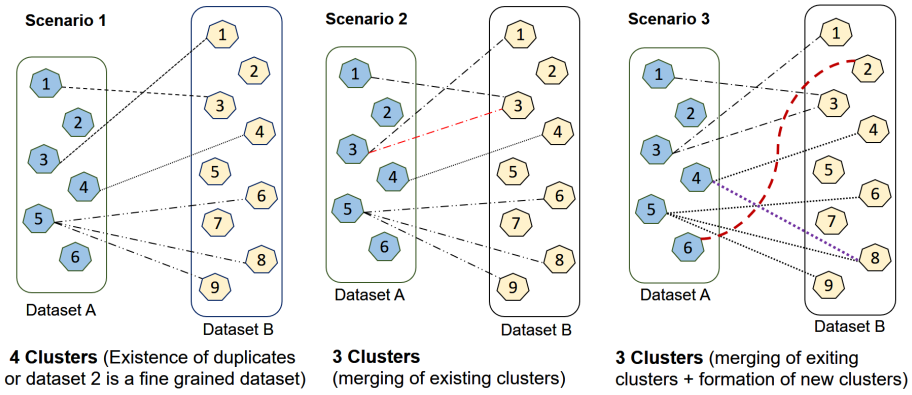


Fig. 6: Decrease/Increase of ILNs

To understand how link-networks are formed as we increase the geo-similarity distance, Figure 6 illustrates how ILNs may evolve as we move from strict constraints (scenario 1) to liberal constraints (scenario 2). First, in **scenario 1**, four ILNs are derived from the six links: $c_1 = \{\{a_1\},\{b_3\}\}$, $c_2 = \{\{a_3\},\{b_1\}\}$,

$c_3 = \{\{a_4\}, \{b_4\}\}$ and $c_4 = \{\{a_5\}, \{b_6, b_8, b_9\}\}$. Then, the new link between a_3 and b_3 in **scenario 2** forces c_1 and c_2 to ***merge***. We now have a total of three ILNs: $c_1 = \{\{a_1, a_3\}, \{b_1, b_3\}\}$, $c_3 = \{\{a_4\}, \{b_4\}\}$ and $c_4 = \{\{a_5\}, \{b_6, b_8, b_9\}\}$. Finally, in **scenario 3**, two new links appear. The first link between a_4 and b_8 causes the merging of c_3 and c_4 while the second link connecting a_6 to b_2 causes the creation of a new ILN. Thereby, the total number of ILNs remains 3. The scenarios depicted in Figure 6 show that, as the ILN constraints become more liberal, the number of links discovered increases while the number of ILNs may increase, remain equal, or even decrease. In other words, when the matching conditions become liberal or more strict, two types of event are likely to happen: (1) formation of new ILNs and/or (2) merging of ILNs. Table 5, shows that, in experiment 2, phenomenon (1) overtakes (2), which explains the increase in the number of ILNs as the near-by distance increases.

6.3 Result and Analysis

Overall, as illustrated in Table 5, the number of ILNs generated in this experiment increased with the increase of the geo-similarity radius. Within a radius of 50 meters, a total of 230 ILNs are generated based on geo-distance only. This number reached 841 ILNs at a 2 kilometres radius. After performing name matching, many links are pruned. Depending on the matching radius, the number of ILNs then varies from 36 to 371.

Statistics on ILNs of size > 2						
	50 meters		**500 meters**		**2 kilometres**	
Size	geo-only	geo+names	geo-only	geo+names	geo-only	geo+names
≥ 3	230	36	738	168	841	371

Table 5: link-network overview.

Due to manpower limitations we concentrate our efforts evaluating networks of size 3. These ILNs cover about 86% of the overall ILNs within 50m radius and 92% within 500m and 2k radius.

Table 6 shows the results of pitching our e_Q metric against the human evaluation of the identity-networks under both the geo-only and the geo+names conditions. The confusion matrices depicted in Table 7 and Table 8 detail the machine quality judgements versus human evaluations of the networks generated within 2 kilometres radius under both conditions. We show the 2km radius as example since it this is likely to generate the worst quality ILNs, providing a severe test for our e_Q metric.

Analysis. In this experiment, we test the behaviour of the proposed e_Q metric in both noisy (*proximity only*) and noise-less (*proximity plus name*) scenarios. The increase of the proximity distance combined with the name refinement positively affects the number of valid candidates, even though it does not fully help in achieving completeness. More importantly for validating our hypothesis, depending on the *reliability of the identity-criteria*, the proposed e_Q metric is in

	50 meters		500 meters		2 kilometres	
Size	geo-only	geo+names	geo-only	geo+names	geo-only	geo+names
Machine statistics on ILN's of size 3						
$= 3$	92	31	249	155	198	342
Machine	M_{good}: 45 M_{maybe}: 0 M_{bad}: 47	M_{good}: 19 M_{maybe}: 12 M_{bad}: 0	M_{good}: 115 M_{maybe}: 0 M_{bad}: 134	M_{good}: 127 M_{maybe}: 0 M_{bad}: 28	M_{good}: 81 M_{maybe}: 0 M_{bad}: 117	M_{good}: 279 M_{maybe}: 0 M_{bad}: 63
Human evaluation on ILN's of size 3						
$= 3$	$\mathbf{F_1 = 0.693}$	$\mathbf{F_1 = 0.826}$	$\mathbf{F_1 = 0.682}$	$\mathbf{F_1 = 0.909}$	$\mathbf{F_1 = 0.803}$	$\mathbf{F_1 = 0.912}$
Human	H_{good}: 31 H_{maybe}:4 H_{bad}: 57	H_{good}: 27 H_{maybe}:1 H_{bad}: 3	H_{good}: 64 H_{maybe}:7 H_{bad}: 176	H_{good}: 148 H_{maybe}:1 H_{bad}: 6	H_{good}: 61 H_{maybe}:3 H_{bad}: 134	H_{good}: 322 H_{maybe}:1 H_{bad}: 3

Table 6: Automated flagging versus human evaluation.

		198 GROUND TRUTHS			
		GT. Pos. 61	GT. neg. 137		
PREDICT	POSITIVE 81	True Pos. 57	False Pos. 24	Precision 0.704	False Discovery Rate 0.296
	NEGATIVE 117	False Neg. 4	True Neg. 113	F. Omission Rate 0.034	Neg. Predictive Value 0.966
		Recall 0.934	Fall-out 0.175	Positive L. Ratio 4.021	F1 score — Accuracy 0.803 — 0.859

Table 7: Confusion matrix for IDLINEs of size 3, 2km, geo-only.

		342 GROUND TRUTHS			
		GT. Pos. 322	GT. neg. 20		
PREDICT	POSITIVE 279	True Pos. 274	False Pos. 5	Precision 0.982	False Discovery Rate 0.018
	NEGATIVE 63	False Neg. 48	True Neg. 15	F. Omission Rate 0.762	Neg. Predictive Value 0.238
		Recall 0.851	Fall-out 0.25	Positive. L. Ratio 3.928	F1 score — Accuracy 0.912 — 0.845

Table 8: Confusion matrix for ILNs of size 3, 2km, geo+names

general able to exclude potentially less good networks in a noisy environment and to include more potentially good networks in noise-less environment. In addition, on the one hand, the relatively low F_1 measures displayed in Table 9 in noisy scenarios, highlight that for the data at hand, proximity alone is not a good enough criterion for identity. On the other hand, the relatively high F_1 measures in noise-less scenarios is an indication of stability and consistency that is in line with results outlined in experiment 1.

The results depicted in Table 9 show an uneven distribution of the candidate-sets. In a relatively balanced candidate-set scenario, our approach works well as can be seen in the first experiment and in the *proximity only* scenario. However, even though in extreme cases (*proximity plus name*) the Majority Class Classifier takes the lead, the network metric does not fall far behind.

Majority Class Classifier vs Network Metrics $\frac{MajorityClassClassifier}{NetworkMetrics}$ GT = Ground Truth GT_P = Ground Truth Positive GT_N = Ground Truth Negative				
50m Before	GT=92	GT_P=30 GT_N=62	F_1 : $\frac{-}{\mathbf{0.693}}$	ACC: $\frac{0.674}{\mathbf{0.75}}$	NPV: $\frac{0.674}{\mathbf{0.915}}$
500m Before	GT=249	GT_P=66 GT_N=183	F_1 : $\frac{-}{\mathbf{0.696}}$	ACC: $\frac{0.735}{\mathbf{0.779}}$	NPV: $\frac{0.735}{\mathbf{0.978}}$
2km Before	GT=198	GT_P=61 GT_N=137	F_1 : $\frac{-}{\mathbf{0.803}}$	ACC: $\frac{0.692}{\mathbf{0.859}}$	NPV: $\frac{0.692}{\mathbf{0.966}}$
50m After	GT=31	GT_P=27 GT_N=4	F_1 : $\frac{\mathbf{0.931}}{0.826}$	ACC: $\frac{\mathbf{0.871}}{0.742}$	NPV: $\frac{-}{\mathbf{0.333}}$
500m After	GT=155	GT_P=148 GT_N=7	F_1 : $\frac{\mathbf{0.977}}{0.909}$	ACC: $\frac{\mathbf{0.955}}{0.839}$	NPV: $\frac{-}{\mathbf{0.179}}$
2km After	GT=342	GT_P=322 GT_N=20	F_1 : $\frac{\mathbf{0.97}}{0.912}$	ACC: $\frac{\mathbf{0.942}}{0.845}$	NPV: $\frac{-}{\mathbf{0.238}}$

Table 9: Network-metric result versus the MCC baseline: Expert sample.

7 Conclusion and further work

7.1 Conclusion

Entity linking is an essential step in the use of multiple datasets on the semantic web. Since entity linking algorithms have less than perfect precision, the links found by such algorithms must often be validated by users. Since this is both an costly and an error-prone process, it is desirable to have computer support that can accurately estimate the quality of links between entities.

In this paper, we have proposed a metric for precisely this purpose: it estimates the quality of links between entities from multiple (> 2) datasets, using a combination of metrics based on the network formed by these links. Our metric captures the intuition that high redundancy in such a linking-network correlates with high quality.

We have tested our metric in two different scenarios, using a collection of 6 widely used social science datasets. In each of our experimental settings, we compared the predictions of link quality by our metric against human judgements on hundreds of networks involving thousands of links. In each evaluation, our metric correlated strongly with human judgement ($0.806 \leq F_1 \leq 1$), and it consistently beats a Majority Class Classifier baseline (except in cases where this is numerically near impossible because of a highly skewed class distribution). In the experimental condition where we deliberatly constructed very noisy link-networks, we showed that our metric is in general able to exclude potentially less good networks in a noisy environment and to include more potentially good networks in noise-less environment.

To the best of our knowledge, this is the first work that uses simple network metrics to successfully estimate the quality of links between entities from more than 2 datasets.

7.2 Future work

Including link strength. Our metric e_Q is based on the presence and absence of links, but does not consider any strength associated with these links. However, many practical entity linking algorithms do in fact produce some kind of confidence score for each link that they find. We are currently working on refinements of e_Q that use such scores to estimate the accuracy with which e_Q makes its predictions.

Dynamic link adjustment. The current work simply takes the output of an external linking algorithm as given, and tries to estimate the quality of that output. A closer coupling between our metric and a linking algorithm would allow the linking algorithm to dynamically adjust its output based on the quality estimates that are provided by our metric. Similarly, the current metric could be embedded in a user-interface where the user gives the final judgement to accept or reject an ILN, taking the score of our metric into account. The weights of the links (as mentioned in the previous item) would also help to guide the user's decision.

References

1. O. Hassanzadeh, A. Kementsietsidis, L. Lim, R. J. Miller, and M. Wang. A framework for semantic link discovery over relational data. In *18th ACM conference on Information and knowledge management*, pages 1027–1036. ACM, 2009.
2. O. Hassanzadeh, R. Xin, R. J. Miller, A. Kementsietsidis, L. Lim, and M. Wang. Linkage query writer. *Proceedings of the VLDB Endowment*, 2(2):1590–1593, 2009.
3. R. Usbeck, A.-C. N. Ngomo, M. Röder, D. Gerber, S. A. Coelho, S. Auer, and A. Both. AGDISTIS-graph-based disambiguation of named entities using linked data. In *The Semantic Web–ISWC 2014*, pages 457–471. Springer, 2014.
4. J. Volz, C. Bizer, M. Gaedke, and G. Kobilarov. Discovering and maintaining links on the web of data. In *ISWC*, pages 650–665. Springer, 2009.

Appendix on all confusion matrices supporting the analysis

Name Similarity Cluster of Size 5

		391 GROUND TRUTHS			
	*** BASELINE ***	GT. Pos. 272	GT. neg. 119		
PREDICT	POSITIVE 391	True Pos. 272	False Pos. 119	Precision 0.696	False Discovery Rate 0.304
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.696	F1 score — Accuracy 0.821 — 0.696

Table 10: Confusion matrix for link-networks of size 5 (MCC).

		391 GROUND TRUTHS			
		GT. Pos. 272	GT. neg. 119		
PREDICT	POSITIVE 289	True Pos. 231	False Pos. 58	Precision 0.799	False Discovery Rate 0.201
	NEGATIVE 102	False Neg. 41	True Neg. 61	F. Omission Rate 0.402	Neg. Predictive Value 0.598
		Recall 0.849	Fall-out 0.487	Positive L. Ratio 1.641	F1 score — Accuracy 0.824 — 0.747

Table 11: Confusion matrix for link-networks of size 5 (Evaluation).

Name Similarity Cluster of Size 6

		224 GROUND TRUTHS			
	*** BASELINE ***	GT. Pos. 139	GT. neg. 85		
PREDICT	POSITIVE 224	True Pos. 139	False Pos. 85	Precision 0.621	False Discovery Rate 0.379
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.621	F1 score — Accuracy 0.766 — 0.621

Table 12: Confusion matrix for link-networks of size 6 (MCC).

		224 GROUND TRUTHS			
		GT. Pos. 139	GT. neg. 85		
PREDICT	POSITIVE 145	True Pos. 116	False Pos. 29	Precision 0.8	False Discovery Rate 0.2
	NEGATIVE 79	False Neg. 23	True Neg. 56	F. Omission Rate 0.291	Neg. Predictive Value 0.709
		Recall 0.835	Fall-out 0.341	Positive L. Ratio 2.346	F1 score — Accuracy 0.817 — 0.768

Table 13: Confusion matrix for link-networks of size 6 (Evaluation).

Name Similarity Cluster of Size 7

		96 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 50	GT. neg. 46		
PREDICT	POSITIVE 96	True Pos. 50	False Pos. 46	Precision 0.521	False Discovery Rate 0.479
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.521	F1 score — Accuracy 0.685 — 0.521

Table 14: Confusion matrix for link-networks of size 7 (MCC).

		96 GROUND TRUTHS			
		GT. Pos. 50	GT. neg. 46		
PREDICT	POSITIVE 54	True Pos. 42	False Pos. 12	Precision 0.778	False Discovery Rate 0.222
	NEGATIVE 42	False Neg. 8	True Neg. 34	F. Omission Rate 0.19	Neg. Predictive Value 0.81
		Recall 0.84	Fall-out 0.261	Positive L. Ratio 2.98	F1 score — Accuracy 0.808 — 0.792

Table 15: Confusion matrix for link-networks of size 7 (Evaluation).

Name Similarity Cluster of Size 8

		66 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 35	GT. neg. 31		
PREDICT	POSITIVE 66	True Pos. 35	False Pos. 31	Precision 0.53	False Discovery Rate 0.47
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.53	F1 score — Accuracy 0.693 — 0.53

Table 16: Confusion matrix for link-networks of size 8 (MCC).

		66 GROUND TRUTHS			
		GT. Pos. 35	GT. neg. 31		
PREDICT	POSITIVE 32	True Pos. 27	False Pos. 5	Precision 0.844	False Discovery Rate 0.156
	NEGATIVE 34	False Neg. 8	True Neg. 26	F. Omission Rate 0.235	Neg. Predictive Value 0.765
		Recall 0.771	Fall-out 0.161	Positive L. Ratio 5.241	F1 score — Accuracy 0.806 — 0.803

Table 17: Confusion matrix for link-networks of size 8 (Evaluation).

Name Similarity Cluster of Size 9

		45 GROUND TRUTHS			
	*** BASELINE ***	GT. Pos. 21	GT. neg. 24		
PREDICT	POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 45	False Neg. 21	True Neg. 24	F. Omission Rate 0.467	Neg. Predictive Value 0.533
		Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.533

Table 18: Confusion matrix for link-networks of size 9 (MCC).

		45 GROUND TRUTHS			
		GT. Pos. 21	GT. neg. 24		
PREDICT	POSITIVE 26	True Pos. 21	False Pos. 5	Precision 0.808	False Discovery Rate 0.192
	NEGATIVE 19	False Neg. 0	True Neg. 19	F. Omission Rate 0.0	Neg. Predictive Value 1.0
		Recall 1.0	Fall-out 0.208	Positive L. Ratio 3.883	F1 score — Accuracy 0.894 — 0.889

Table 19: Confusion matrix for link-networks of size 9 (Evaluation).

Name Similarity Cluster of Size 10

		24 GROUND TRUTHS			
	*** BASELINE ***	GT. Pos. 8	GT. neg. 16		
PREDICT	POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 24	False Neg. 8	True Neg. 16	F. Omission Rate 0.333	Neg. Predictive Value 0.667
		Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.667

Table 20: Confusion matrix for link-networks of size 10 (MCC).

		24 GROUND TRUTHS			
		GT. Pos. 8	GT. neg. 16		
PREDICT	POSITIVE 7	True Pos. 7	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 17	False Neg. 1	True Neg. 16	F. Omission Rate 0.059	Neg. Predictive Value 0.941
		Recall 0.875	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy 0.933 — 0.958

Table 21: Confusion matrix for link-networks of size 10 (Evaluation).

Name Similarity Cluster of Size 3: MCC vs. Expert

		64 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 56	GT. neg. 8		
PREDICT	POSITIVE 64	True Pos. 56	False Pos. 8	Precision 0.875	False Discovery Rate 0.125
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.875	F1 score — Accuracy 0.933 — 0.875

Table 22: Confusion matrix for link-networks of size 3 (MCC).

		64 GROUND TRUTHS			
		GT. Pos. 56	GT. neg. 8		
PREDICT	POSITIVE 60	True Pos. 54	False Pos. 6	Precision 0.9	False Discovery Rate 0.1
	NEGATIVE 4	False Neg. 2	True Neg. 2	F. Omission Rate 0.5	Neg. Predictive Value 0.5
		Recall 0.964	Fall-out 0.75	Positive L. Ratio 1.2	F1 score — Accuracy 0.931 — 0.875

Table 23: Confusion matrix for link-networks of size 3 (Expert).

Name Similarity Cluster of Size 4: MCC vs. Expert

		24 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 19	GT. neg. 5		
PREDICT	POSITIVE 24	True Pos. 19	False Pos. 5	Precision 0.792	False Discovery Rate 0.208
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.792	F1 score — Accuracy 0.884 — 0.792

Table 24: Confusion matrix for link-networks of size 4 (MCC).

		24 GROUND TRUTHS			
		GT. Pos. 19	GT. neg. 5		
PREDICT	POSITIVE 22	True Pos. 18	False Pos. 4	Precision 0.818	False Discovery Rate 0.182
	NEGATIVE 2	False Neg. 1	True Neg. 1	F. Omission Rate 0.5	Neg. Predictive Value 0.5
		Recall 0.947	Fall-out 0.8	Positive L. Ratio 1.023	F1 score — Accuracy 0.878 — 0.792

Table 25: Confusion matrix for link-networks of size 4 (Expert).

Name Similarity Cluster of Size 5: MCC vs. Expert

		15 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 14	GT. neg. 1		
PREDICT	POSITIVE 15	True Pos. 14	False Pos. 1	Precision 0.933	False Discovery Rate 0.067
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.933	F1 score — Accuracy 0.966 — 0.933

Table 26: Confusion matrix for link-networks of size 5 (MCC).

		15 GROUND TRUTHS			
		GT. Pos. 14	GT. neg. 1		
PREDICT	POSITIVE 14	True Pos. 13	False Pos. 1	Precision 0.929	False Discovery Rate 0.071
	NEGATIVE 1	False Neg. 1	True Neg. 0	F. Omission Rate 1.0	Neg. Predictive Value 0.0
		Recall 0.929	Fall-out 1.0	Positive L. Ratio 0.929	F1 score — Accuracy 0.929 — 0.867

Table 27: Confusion matrix for link-networks of size 5 (Expert).

Name Similarity Cluster of Size 6: MCC vs. Expert

		19 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 14	GT. neg. 5		
PREDICT	POSITIVE 19	True Pos. 14	False Pos. 5	Precision 0.737	False Discovery Rate 0.263
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.737	F1 score — Accuracy 0.848 — 0.737

Table 28: Confusion matrix for link-networks of size 6 (MCC).

		19 GROUND TRUTHS			
		GT. Pos. 14	GT. neg. 5		
PREDICT	POSITIVE 19	True Pos. 14	False Pos. 5	Precision 0.737	False Discovery Rate 0.263
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.737	F1 score — Accuracy 0.848 — 0.737

Table 29: Confusion matrix for link-networks of size 6 (Expert).

Name Similarity Cluster of Size 7: MCC vs. Expert

		12 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 10	GT. neg. 2		
PREDICT	POSITIVE 12	True Pos. 10	False Pos. 2	Precision 0.833	False Discovery Rate 0.167
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.833	F1 score — Accuracy 0.909 — 0.833

Table 30: Confusion matrix for link-networks of size 7 (MCC).

		12 GROUND TRUTHS			
		GT. Pos. 10	GT. neg. 2		
PREDICT	POSITIVE 10	True Pos. 10	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 2	False Neg. 0	True Neg. 2	F. Omission Rate 0.0	Neg. Predictive Value 1.0
		Recall 1.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 31: Confusion matrix for link-networks of size 7 (Expert).

Name Similarity Cluster of Size 8: MCC vs. Expert

		4 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 4	GT. neg. 0		
PREDICT	POSITIVE 4	True Pos. 4	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out -	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 32: Confusion matrix for link-networks of size 8 (MCC).

		4 GROUND TRUTHS			
		GT. Pos. 4	GT. neg. 0		
PREDICT	POSITIVE 4	True Pos. 4	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out -	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 33: Confusion matrix for link-networks of size 8 (Expert).

Name Similarity Cluster of Size 9: MCC vs. Expert

		9 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 8	GT. neg. 1		
PREDICT	POSITIVE 9	True Pos. 8	False Pos. 1	Precision 0.889	False Discovery Rate 0.111
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.889	F1 score — Accuracy 0.941 — 0.889

Table 34: Confusion matrix for link-networks of size 9 (MCC).

		9 GROUND TRUTHS			
		GT. Pos. 8	GT. neg. 1		
PREDICT	POSITIVE 8	True Pos. 8	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 1	False Neg. 0	True Neg. 1	F. Omission Rate 0.0	Neg. Predictive Value 1.0
		Recall 1.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 35: Confusion matrix for link-networks of size 9 (Expert).

Name Similarity Cluster of Size 10: MCC vs. Expert

		1 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 1	GT. neg. 0		
PREDICT	POSITIVE 1	True Pos. 1	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out -	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 36: Confusion matrix for link-networks of size 10 (MCC).

		1 GROUND TRUTHS			
		GT. Pos. 1	GT. neg. 0		
PREDICT	POSITIVE 1	True Pos. 1	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out -	Positive L. Ratio -	F1 score — Accuracy 1.0 — 1.0

Table 37: Confusion matrix for link-networks of size 10 (Expert).

50 meters proximity

		92 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 30	GT. neg. 62		
PREDICT	POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 92	False Neg. 30	True Neg. 62	F. Omission Rate 0.326	Neg. Predictive Value 0.674
		Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.674

Table 38: Confusion matrix for link-networks of size 3, 50m before (MCC).

		92 GROUND TRUTHS			
		GT. Pos. 30	GT. neg. 62		
PREDICT	POSITIVE 45	True Pos. 26	False Pos. 19	Precision 0.578	False Discovery Rate 0.422
	NEGATIVE 47	False Neg. 4	True Neg. 43	F. Omission Rate 0.085	Neg. Predictive Value 0.915
		Recall 0.867	Fall-out 0.306	Positive L. Ratio 1.888	F1 score — Accuracy 0.693 — 0.75

Table 39: Confusion matrix for link-networks of size 3, 50m before (Evaluation).

50 meters proximity + name

		31 GROUND TRUTHS		*** BASE	
	*** BASE LINE ***	GT. Pos. 27	GT. neg. 4		
PREDICT	POSITIVE 31	True Pos. 27	False Pos. 4	Precision 0.871	False Discovery Rate 0.129
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.871	F1 score — Accuracy 0.931 — 0.871

Table 40: Confusion matrix for link-networks of size 3, 50m after (MCC).

		31 GROUND TRUTHS			
		GT. Pos. 27	GT. neg. 4		
PREDICT	POSITIVE 19	True Pos. 19	False Pos. 0	Precision 1.0	False Discovery Rate 0.0
	NEGATIVE 12	False Neg. 8	True Neg. 4	F. Omission Rate 0.667	Neg. Predictive Value 0.333
		Recall 0.704	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy 0.826 — 0.742

Table 41: Confusion matrix for link-networks of size 3, 50m after (Evaluation).

500 meters proximity

		249 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 66	GT. neg. 183		
PREDICT	POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 249	False Neg. 66	True Neg. 183	F. Omission Rate 0.265	Neg. Predictive Value 0.735
		Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.735

Table 42: Confusion matrix for link-networks of size 3, 500m before (MCC).

		249 GROUND TRUTHS			
		GT. Pos. 66	GT. neg. 183		
PREDICT	POSITIVE 115	True Pos. 63	False Pos. 52	Precision 0.548	False Discovery Rate 0.452
	NEGATIVE 134	False Neg. 3	True Neg. 131	F. Omission Rate 0.022	Neg. Predictive Value 0.978
		Recall 0.955	Fall-out 0.284	Positive L. Ratio 1.929	F1 score — Accuracy 0.696 — 0.779

Table 43: Confusion matrix for link-networks of size 3, 500m before (Evaluation).

500 meters proximity + name

		155 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 148	GT. neg. 7		
PREDICT	POSITIVE 155	True Pos. 148	False Pos. 7	Precision 0.955	False Discovery Rate 0.045
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.955	F1 score — Accuracy 0.977 — 0.955

Table 44: Confusion matrix for link-networks of size 3, 500m after (MCC).

2000 meters proximity

		198 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 61	GT. neg. 137		
PREDICT	POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 198	False Neg. 61	True Neg. 137	F. Omission Rate 0.308	Neg. Predictive Value 0.692
		Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.692

Table 45: Confusion matrix for link-networks of size 3, 2 km before (MCC).

		198 GROUND TRUTHS			
		GT. Pos. 61	GT. neg. 137		
PREDICT	POSITIVE 81	True Pos. 57	False Pos. 24	Precision 0.704	False Discovery Rate 0.296
	NEGATIVE 117	False Neg. 4	True Neg. 113	F. Omission Rate 0.034	Neg. Predictive Value 0.966
		Recall 0.934	Fall-out 0.175	Positive L. Ratio 4.021	F1 score — Accuracy 0.803 — 0.859

Table 46: Confusion matrix for link-networks of size 3, 2 km before (Evaluation).

2000 meters proximity + name

		342 GROUND TRUTHS			
	*** BASE LINE ***	GT. Pos. 322	GT. neg. 20		
PREDICT	POSITIVE 342	True Pos. 322	False Pos. 20	Precision 0.942	False Discovery Rate 0.058
	NEGATIVE 0	False Neg. 0	True Neg. 0	F. Omission Rate -	Neg. Predictive Value -
		Recall 1.0	Fall-out 1.0	Positive L. Ratio 0.942	F1 score — Accuracy 0.97 — 0.942

Table 47: Confusion matrix for link-networks of size 3, 2 km after (MCC).

		342 GROUND TRUTHS			
		GT. Pos. 322	GT. neg. 20		
PREDICT	POSITIVE 279	True Pos. 274	False Pos. 5	Precision 0.982	False Discovery Rate 0.018
	NEGATIVE 63	False Neg. 48	True Neg. 15	F. Omission Rate 0.762	Neg. Predictive Value 0.238
		Recall 0.851	Fall-out 0.25	Positive L. Ratio 3.928	F1 score — Accuracy 0.912 — 0.845

Table 48: Confusion matrix for link-networks of size 3, 2 km after (Evaluation).