

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

## Disambiguating Orgreg

Al Idrissou<sup>1,2</sup>, Frank van Harmelen<sup>1</sup>, and Peter van den Besselaar<sup>2</sup>

<sup>1</sup> Department of Computer Science, Vrije Universiteit Amsterdam, NL

<sup>2</sup> 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 these 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.

## 2 Identity Link Network

...

## 3 Related work

...

## 4 Network Properties & the Quality of a Link-Network.

...

## 5 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<sup>3</sup>.

**Grid**<sup>4</sup> 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

---

<sup>3</sup> 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.

<sup>4</sup> <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**<sup>5</sup> 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**<sup>6</sup> 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**<sup>7</sup> 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**<sup>8</sup> 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**<sup>9</sup> 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.

## 6 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.

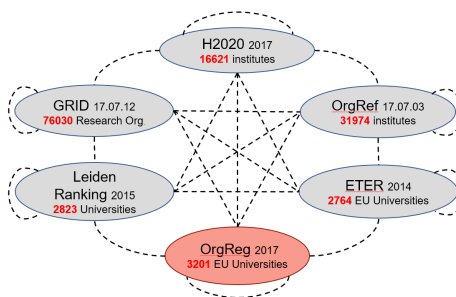


Fig. 1: Disambiguating OrgReg

<sup>5</sup> <http://www.orgref.org>

<sup>6</sup> <http://www.leidenranking.com/>

<sup>7</sup> <https://www.eter-project.com/>

<sup>8</sup> <http://risis.eu/orgreg/>

<sup>9</sup> <http://www.gaeu.com/sv/item/horizon-2020>

## 6.1 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 2).

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.

## 6.2 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 also expect to find ILNs of size  $< 6$ . Figure 2 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  $F_1$  scores. In red, Figure 2 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<sup>10</sup> for each ILN size. As an example<sup>11</sup>, Table 1 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 2 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.

## 6.3 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 3) in which at least one resource is located

<sup>10</sup> [https://en.wikipedia.org/wiki/Confusion\\_matrix](https://en.wikipedia.org/wiki/Confusion_matrix)

<sup>11</sup> All confusion matrices supporting the experiments analysis could be found on the RISIS page at <http://sms.risis.eu/assets/pdf/metrics-link-network.pdf>

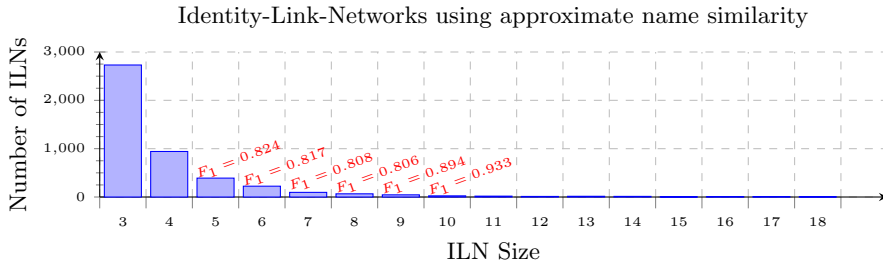


Fig. 2: Clusters extracted from the lens.

		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 1: Confusion matrix for link-networks of size 8.

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

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

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

#### 6.4 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 2 and 0.930 in Table 3). Furthermore,  $e_Q$  outperformed a majority-class classifier in Table 2 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.

## 7 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.

## 7.1 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 were 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.

## 7.2 Strict vs. Liberal Clustering

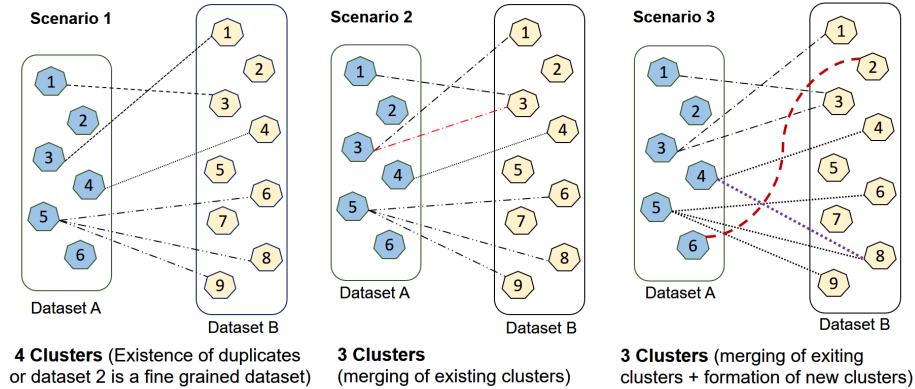


Fig. 3: Decrease/Increase of ILNs

To understand how link-networks are formed as we increase the geo-similarity distance, Figure 3 illustrates how ILNs may evolve as we move from strict constraints (scenario 1) to liberal constraints (scenario 3). 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 3 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 4, shows that, in experiment 2, phenomenon (1) overtakes (2), which explains the increase in the number of ILNs as the near-by distance increases.

### 7.3 Result and Analysis

Overall, as illustrated in Table 4, 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 4: 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 5 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 6 and Table 7 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<sup>12</sup> since 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

<sup>12</sup> All confusion matrices supporting the experiments analysis could be found on the RISIS page at <http://sms.risis.eu/assets/pdf/metrics-link-network.pdf>



		50 meters		500 meters		2 kilometres	
Size	geo-only	geo+names	geo-only	geo+names	geo-only	geo+names	
<b>Machine statistics on ILN's of size 3</b>							
= 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	
<b>Human evaluation on ILN's of size 3</b>							
= 3	<b>F<sub>1</sub> = 0.693</b>	<b>F<sub>1</sub> = 0.826</b>	<b>F<sub>1</sub> = 0.682</b>	<b>F<sub>1</sub> = 0.909</b>	<b>F<sub>1</sub> = 0.803</b>	<b>F<sub>1</sub> = 0.912</b>	
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 5: Automated flagging versus human evaluation.

		198 GROUND TRUTHS			
		GT. Pos. 61	GT. neg. 137		
PREDICT	POSITIVE	True Pos. 57	False Pos. 24	Precision 0.704	False Discovery Rate 0.296
	NEGATIVE	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 6: Confusion matrix for IDLINES of size 3, 2km, geo-only.

		342 GROUND TRUTHS			
		GT. Pos. 322	GT. neg. 20		
PREDICT	POSITIVE	True Pos. 274	False Pos. 5	Precision 0.982	False Discovery Rate 0.018
	NEGATIVE	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 7: Confusion matrix for ILNs of size 3, 2km, geo+names

in achieving completeness. More importantly for validating our hypothesis, depending on the *reliability of the identity-criteria*, the proposed  $e_Q$  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. In addition, on the one hand, the relatively low  $F_1$  measures displayed in Table 8 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 8 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						
<i>MajorityClassClassifier</i>						
<i>NetworkMetrics</i>						
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: $\frac{\mathbf{0.674}}{\mathbf{0.75}}$	NPV: $\frac{\mathbf{0.674}}{\mathbf{0.915}}$
500m Before	GT=249	$GT_P=66$	$GT_N=183$	$F_1 : \frac{-}{\mathbf{0.696}}$	ACC: $\frac{\mathbf{0.735}}{\mathbf{0.779}}$	NPV: $\frac{\mathbf{0.735}}{\mathbf{0.978}}$
2km Before	GT=198	$GT_P=61$	$GT_N=137$	$F_1 : \frac{-}{\mathbf{0.803}}$	ACC: $\frac{\mathbf{0.692}}{\mathbf{0.859}}$	NPV: $\frac{\mathbf{0.692}}{\mathbf{0.966}}$
50m After	GT=31	$GT_P=27$	$GT_N=4$	$F_1 : \frac{\mathbf{0.931}}{\mathbf{0.826}}$	ACC: $\frac{\mathbf{0.871}}{\mathbf{0.742}}$	NPV: $\frac{-}{\mathbf{0.333}}$
500m After	GT=155	$GT_P=148$	$GT_N=7$	$F_1 : \frac{\mathbf{0.977}}{\mathbf{0.909}}$	ACC: $\frac{\mathbf{0.955}}{\mathbf{0.839}}$	NPV: $\frac{-}{\mathbf{0.179}}$
2km After	GT=342	$GT_P=322$	$GT_N=20$	$F_1 : \frac{\mathbf{0.97}}{\mathbf{0.912}}$	ACC: $\frac{\mathbf{0.942}}{\mathbf{0.845}}$	NPV: $\frac{-}{\mathbf{0.238}}$

Table 8: Network-metric result versus the MCC baseline: Expert sample.

Like in the first experiment, for further evaluation, we extracted a sample based on ILNs in which at least one organisation originates from the Netherlands. Out of the **107** sampled ILNs, the Dutch domain expert deviated from the first evaluation in only 1 case, and in favour of the metrics ( $F_1 : \frac{-}{\mathbf{0.711}}$ ,  $ACC : \frac{\mathbf{0.663}}{\mathbf{0.761}}$  and  $NPV : \frac{\mathbf{0.663}}{\mathbf{0.915}}$ ). This is observed in the noisy scenario for the radius of 50m.

## 8 Conclusion and further work

### 8.1 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 a 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 deliberately 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.

## 8.2 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.

**Parameter tuning.** In this work, we empirically determine the 1.6 sigmoid hyper-parameter and the discrete warning intervals. Experimenting on fine-tuning these parameters using the current ground truth data and data from other domains would help understand how and when different choices could lead to an increase or a decrease of the of the metrics' predictive power depending on the use case.

## 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			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		272	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 9: Confusion matrix for link-networks of size 5 (MCC).

		391 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		272	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 10: Confusion matrix for link-networks of size 5 (Evaluation).

### Name Similarity Cluster of Size 6

		224 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		139	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 11: Confusion matrix for link-networks of size 6 (MCC).

		224 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		139	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 12: Confusion matrix for link-networks of size 6 (Evaluation).

## Name Similarity Cluster of Size 7

		96 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		50	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 13: Confusion matrix for link-networks of size 7 (MCC).

		96 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		50	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 14: Confusion matrix for link-networks of size 7 (Evaluation).

## Name Similarity Cluster of Size 8

		66 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		35	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 15: Confusion matrix for link-networks of size 8 (MCC).

		66 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		35	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 16: Confusion matrix for link-networks of size 8 (Evaluation).

## Name Similarity Cluster of Size 9

		45 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		21	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 17: Confusion matrix for link-networks of size 9 (MCC).

		45 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		21	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 18: Confusion matrix for link-networks of size 9 (Evaluation).

## Name Similarity Cluster of Size 10

		24 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		8	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 19: Confusion matrix for link-networks of size 10 (MCC).

		24 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		8	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 20: Confusion matrix for link-networks of size 10 (Evaluation).

### Name Similarity Cluster of Size 3: MCC vs. Expert

		64 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		56	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 21: Confusion matrix for link-networks of size 3 (MCC).

		64 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		56	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 22: Confusion matrix for link-networks of size 3 (Expert).

### Name Similarity Cluster of Size 4: MCC vs. Expert

		24 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		19	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 23: Confusion matrix for link-networks of size 4 (MCC).

		24 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		19	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 24: Confusion matrix for link-networks of size 4 (Expert).

## Name Similarity Cluster of Size 5: MCC vs. Expert

		15 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		14	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 25: Confusion matrix for link-networks of size 5 (MCC).

		15 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		14	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 26: Confusion matrix for link-networks of size 5 (Expert).

## Name Similarity Cluster of Size 6: MCC vs. Expert

		19 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		14	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 27: Confusion matrix for link-networks of size 6 (MCC).

		19 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		14	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 (Expert).



### Name Similarity Cluster of Size 7: MCC vs. Expert

		12 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		10	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 29: Confusion matrix for link-networks of size 7 (MCC).

		12 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		10	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 30: Confusion matrix for link-networks of size 7 (Expert).

### Name Similarity Cluster of Size 8: MCC vs. Expert

		4 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		4	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 31: Confusion matrix for link-networks of size 8 (MCC).

		4 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		4	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 (Expert).

### Name Similarity Cluster of Size 9: MCC vs. Expert

		9 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		8	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 33: Confusion matrix for link-networks of size 9 (MCC).

		9 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		8	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 34: Confusion matrix for link-networks of size 9 (Expert).

### Name Similarity Cluster of Size 10: MCC vs. Expert

		1 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		1	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 35: Confusion matrix for link-networks of size 10 (MCC).

		1 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		1	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 (Expert).

## 50 meters proximity

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

Table 37: Confusion matrix for link-networks of size 3, 50m before (MCC).

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

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

## 50 meters proximity + name

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

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

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

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

50 meters proximity: all remain the same except this

		92 GROUND TRUTHS			
		GT. Pos. 31	GT. neg. 61		
PREDICT	*** BASE LINE *** POSITIVE 0	True Pos. 0	False Pos. 0	Precision -	False Discovery Rate -
	NEGATIVE 92	False Neg. 31	True Neg. 61	F. Omission Rate 0.337	Neg. Predictive Value 0.663
	Recall 0.0	Fall-out 0.0	Positive L. Ratio -	F1 score — Accuracy - — 0.663	

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

		92 GROUND TRUTHS			
		GT. Pos. 31	GT. neg. 61		
PREDICT	POSITIVE 45	True Pos. 27	False Pos. 18	Precision 0.6	False Discovery Rate 0.4
	NEGATIVE 47	False Neg. 4	True Neg. 43	F. Omission Rate 0.085	Neg. Predictive Value 0.915
	Recall 0.871	Fall-out 0.295	Positive L. Ratio 2.034	F1 score — Accuracy 0.711 — 0.761	

Table 42: Confusion matrix for link-networks of size 3, 50m before (Expert).

## 500 meters proximity

		249 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		66	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 43: Confusion matrix for link-networks of size 3, 500m before (MCC).

		249 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		66	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 44: Confusion matrix for link-networks of size 3, 500m before (Evaluation).

## 500 meters proximity + name

		155 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		148	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 45: Confusion matrix for link-networks of size 3, 500m after (MCC).

		155 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		148	7		
PREDICT	POSITIVE 127	True Pos. 125	False Pos. 2	Precision 0.984	False Discovery Rate 0.016
	NEGATIVE 28	False Neg. 23	True Neg. 5	F. Omission Rate 0.821	Neg. Predictive Value 0.179
		Recall 0.845	Fall-out 0.286	Positive L. Ratio 3.441	F1 score — Accuracy 0.909 — 0.839

Table 46: Confusion matrix for link-networks of size 3, 500m after (Evaluation).

## 2000 meters proximity

		198 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		61	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 47: Confusion matrix for link-networks of size 3, 2 km before (MCC).

		198 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		61	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 48: Confusion matrix for link-networks of size 3, 2 km before (Evaluation).

## 2000 meters proximity + name

		342 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		322	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 49: Confusion matrix for link-networks of size 3, 2 km after (MCC).

		342 GROUND TRUTHS			
		GT. Pos.	GT. neg.		
*** BASE LINE ***		322	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 50: Confusion matrix for link-networks of size 3, 2 km after (Evaluation).