Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets[1]

Al Idrissou [a,*], Frankvan Harmelen [a] and Peter van den Besselaar [b]
[a] *Department of Computer Science, Vrije Universiteit Amsterdam, The Netherlands*
E-mails: o.a.k.idrissou@vu.nl, frank.van.harmelen@vu.nl
[b] *Department of Organization Sciences, Vrije Universiteit Amsterdam, The Netherlands*
E-mail: p.a.a.vanden.besselaar@vu.nl

Abstract. Matching entities between datasets is a crucial step for combining multiple datasets on the semantic web. A rich literature exists on different approaches to this entity resolution problem. However, much less work has been done on how to *assess* the quality of such entity links once they have been generated. Evaluation methods for link quality are typically limited to either comparison with a *ground truth dataset* (which is often not available), *manual work* (which is cumbersome and prone to error), or *crowd sourcing* (which is not always feasible, especially if expert knowledge 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 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 can reliably predict their quality. We verify our results in a large experimental study using six datasets from the domain of science, technology and innovation studies, for which we created a gold standard. This gold standard, available online, is an additional contribution of this paper. In addition, we evaluate our metric on a recently published gold standard to confirm our findings.

Keywords: entity resolution, data integration, network metrics

1. Introduction

Matching entities between datasets (known as entity resolution) is a crucial step for the use of multiple datasets on the semantic web. There exists a fair amount of entity resolution tools for *generating* links between pairs of resources: AGDISTIS[2], LIMES[3] Linkage Query Writer [4, 5], SILK [6], etc. However, much fewer methods exist for *validating* the links produced by these methods. Currently, only three validation options are available for such validation: (1) *ground truth*, which is often not available; (2) *manual work*, which is a cumbersome task prone to error; (3) *crowd sourcing*, which is not always feasible especially if specialist knowledge is required. Furthermore, the problem of link evaluation is greatly exacerbated for entity resolution between more than two datasets, because the number of possible links grows rapidly with the number of datasets. Therefore, it is important to investigate ***the accurate automated evaluation of discovered links***. Any answer to this question should generalise beyond the setting of just two datasets, and be applicable to the general setting of links between multiple datasets. In such a multi-dataset scenario, linked resources cluster in small groups that we call *Identity Link Networks (ILNs)*. The goal of this paper is not to propose any new method for entity resolution but instead to provide a method to estimate the quality of an identity link network, and consequently validate a set of discovered links. To do so, ***we hypothesise that the structure of an identity link network correlates with its quality***.

[1] This is an extended version, by invitation, of a paper accepted at the 21st International Conference on Knowledge Engineering and Knowledge Management (EKAW 2018) [1]

[*] Corresponding author. E-mail: o.a.k.idrissou@vu.nl.

We test our hypothesis in two experiments where we show that the proposed metrics indeed reliably estimates the quality of an identity network. We also test our hypothesis on recently published experimental data from ESWC 2018 (see Section 8). Here too, the results confirm that our quality metric reliably predicts human assessment of entity links.

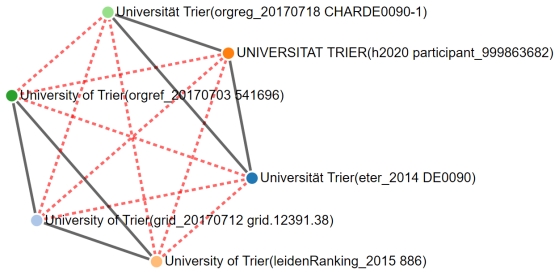
In summary, our contributions is a method that estimates the quality of an identity network of size three or bigger. It is tested against human judgement in three large experiments and correctly classifies large amount of ILNs made available online.[2]

This paper begins with a short motivation in Section 2. Section 3 discusses the related work and section 4 describes the proposed metric. In Section 5 we describe the datasets involved in our experiments. Sections 6, 7 and 8 describe our three experiments. While Section 9 presents possible refinement of the proposed metric, Section 10 evaluates them. Section 11 concludes.

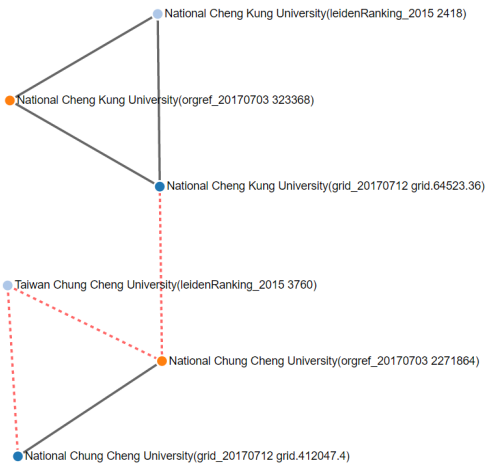
2. Identity Link Networks

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 resolution 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 entity resolution algorithms to represent the same real world entity. An ILN can be derived directly from entity resolution results (Sections 6 and 7), or it may be generated by sophisticated clustering methods as in our experiment in Section 8. In this work we do *not* propose any new entity resolution algorithm. Instead, we propose a method to automatically *evaluate* discovered links, particularly when they involve more than two datasets. Unfortunately, gold standards in initiatives such as OAEI[3] do not go beyond two datasets.

Fig. 1 shows two examples of such ILNs that have been generated by an entity resolution algorithm between entities from six datasets taken from the field of Science, Technology and Innovation studies (STI) (more details in section 5). Fig. **??** 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 the links in the network they form.*



(a) The university of Trier in an ILN across six datasets. The more an ILN resembles a fully connected graph, the more evidence is available to support its identity links.



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

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

Simple Clustering Algorithm. Our aim is not clustering, it is instead the quality approximation of ILNs. So, for reproducibility purposes we present here the straight forward simple clustering algorithm (see algorithm 1) implemented for clustering candidate linked resources. For the purpose of cluster quality estimation, the algorithm also documents the discovered links and their respective strength(s). All basic operations such as **SEARCH** (search for the cluster to with a node belongs to), **INSERTION** (add a node to the set of nodes of a particular cluster, add a link to the set of links of a particular cluster, add a strength to the mapping link → strength of a particular cluster), and **DELETION** (deleting

[2] https://github.com/alkoudouss/Identity-Link-Network-Metric
[3] http://oaei.ontologymatching.org/

a cluster, reassign a cluster to a node) are supported by hash tables $(O(1))$ and therefore minimising to algorithm time complexity ($O(m)$ where m is the size of the input or the number of links to be more precise).

Algorithm 1: Simple resource clustering algorithm & network documentation. All search, insertion and deletion within a cluster is supported with hash tables which allows for a complexity $O(1)$ leading the algorithm to be of $O(m)$ where m is the input size.

input : L, set of tuples $\langle (n_1,\ n_2),\ s\rangle$ representing the mapping $N \times N \to \mathbb{R}$ where $n_i \in N$ and links $(n_1,\ n_2)$ have a strength $s \in \mathbb{R}$

output: $\Gamma \subset \mathcal{P}(N)$, set of cluster-sets C where for each $n_i, n_j \in C,\ n_i =_\Pi n_j$ given a set of criteria Π

begin

- $\Gamma \leftarrow \emptyset$
- **for** $\langle (n_1,\ n_2),\ s\rangle \in L$ **do** `/* O(m) */`
  - **if** $n_1, n_2 \notin C_i$ *for all* $C_i \in \Gamma$ **then**
    - C.(nodes, links, strengths) $\leftarrow \Big(\{n_1, n_2\}, \{(n_1,\ n_2)\}, \{\big((n_1,\ n_2),\ [s]\big)\}\Big)$
    - Γ.add(C)
  - **else if** $n_1 \in C_1 \in \Gamma$ *and* $n_2 \notin C_i$ *for all* $C_i \in \Gamma$ **then**
    - C_1.add($(\{n_2\}, \{(n_1,\ n_2)\}, \{\big((n_1,\ n_2),\ [s]\big)\})$)
  - **else if** $n_2 \in C_2 \in \Gamma$ *and* $n_1 \notin C_i$ *for all* $C_i \in \Gamma$ **then**
    - C_2.add($(\{n_1\}, \{(n_1,\ n_2)\}, \{\big((n_1,\ n_2),\ [s]\big)\})$)
  - **else if** $n_1 \in C_1 \in \Gamma$ *and* $n_2 \in C_2 \in \Gamma$ **then**
    - **if** $C_1 \neq C_2$ **then** `/* they belongs to different clusters */`
      - $C_s \leftarrow$ smallest of C_1 and C_2
      - $C_b \leftarrow$ biggest of C_1 and C_2
      - C_b.add(C_s.items())
      - Γ.delete(C_s)
      - C_b.links.add($(n_1,\ n_2)$)
      - C_b.strength.add($\big((n_1,\ n_2),\ [s]\big)$)
    - **else** `/* they belong to the same cluster */`
      - **if** $(n_1,\ n_2) \in C_1.links$ **then**
        - C_1.strengths[$(n_1,\ n_2)$].add(s)
      - **else**
        - C_1.links.add($(n_1,\ n_2)$)
        - C_1.strength.add($\big((n_1,\ n_2),\ [s]\big)$)
- **return** Γ

3. Related work

We briefly discuss a number of related areas from the literature, and indicate how our work differs from these in aim and scope.

Schema matching. Much work in the literature focuses on ontology matching, especially schema matching [7]. Some rely on concept distance or an extended version of it [8–10]. Some rely on alignment similarities [11], others relies on formal logical conflicts between ontologies to detect and possibly repair mappings at a schema-level [12]. The current paper does not aim to match ontologies, nor does it critically rely on using ontological or schema information. We only assume the existence of external entity resolution algorithms for suggesting links between entities. Such algorithms may or may not exploit ontological information, but this does not affect our central hypothesis.

Information gain. The work in [13] also uses network structure to evaluate link quality, but in a very different way. The main intuition there is that an individual link in an ILN is more reliable when it leads to a greater information gain. The paper does not consider the structure of the ILN as a whole, as we do in this paper.

Entity clustering. Part of the literature also uses clustering of the digital representations of the same real world entity in one or multiple sources. While their data sources are mainly unstructured [14, 15], our interest lies in clusters derived from the mappings of entities exclusively across knowledge-bases. In addition, they also do not consider the structure of the ILN as a whole. Another part of the literature specifically focuses on clustering algorithms. The FAMER [16] framework for example provides and compares seven different link-based entity clustering approaches. The aim of our work is different from all of these. Whereas these works use clustering algorithms to *construct* entity resolutions, we show how a cluster-based metric can be used to *assess* the quality of a network of entity links, irrespective of how these links were generated.

Network metrics. The work by Guéret et al. [17] is one of the few papers to our knowledge that uses network metrics to assess the quality of links. The key point that separates this work from ours is that it uses *local* network features, i.e. only the direct neighbours of a single node, while we employ *global* network features. [18] also addresses the same challenge. It evaluates a given cluster G by comparing it to a reference cluster R based on the number of splits and merges required to go from G to R. Our proposed metric does not need such a reference cluster, and is hence more easily applicable.

4. Network Properties & Quality of a Link-Network

Figure 2 illustrates a set of six simple network topologies over the same number of nodes. Our proposed metric 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 establish an identity-link between two resources (again, not the case for any other topology). Hence, intuitively, the amount of redundancy between paths in an ILN is an indicator for the quality of the links in the ILN. We will capture these and similar intuitions using three different global graph features over ILNs: *Bridge*, *Diameter* and *Closure*.

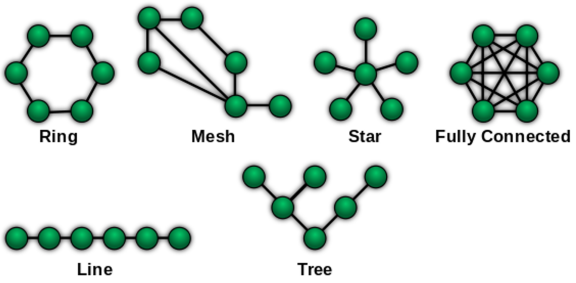


Fig. 2. Example of network topologies.
Source: https://en.wikipedia.org/wiki/Network_topology

We will now first define and explain the rationale behind each metric, then normalise each metric to values[4] between 0 and 1, and finally average the sum of all metrics to obtain the metric which we will use for estimating the quality of the ILN.

Bridge Metric. A bridge (also known as an isthmus or a cut-edge) in a graph is an edge whose removal increases the number of connected components of the graph, or equivalently, an edge that does not belong to any cycle. The intuition for this measure is that a bridge in an ILN suggests a potentially problematic link which is not corroborated by any other links. As a graph with n nodes contains at most $n-1$ bridges (e.g. in a ***Line*** network), the bridge value is normalised as $n_b = \frac{B}{n-1}$, 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 constant), we "soften" the value of n_b with a sigmoid function: $n'_b = max(n_b, sigmoid(B))$, where the 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.

Diameter Metric. 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. In a fully connected graph of n nodes, the diameter $D = 1$. The longest diameter is observed in a ***Line*** network structure, with $D = n-1$ for a line network of n nodes. To scale to the [0,1] interval, the diameter is normalised as $n_d = \frac{D-1}{(n-1)-1}$. 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 Metric. In a connected graph of n nodes, the closure is the ratio of the number of arcs A in the graph over the total number of possible arcs $\frac{1}{2}n(n-1)$. 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 metric as $n_c = 1 - \frac{A}{\frac{1}{2}n(n-1)}$. The minimum number of connections is $n-1$, as observed in ***Line*** and ***Star*** network structures.

Estimated Quality Metric. 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: $n_c = 0$ or $n'_d = 0$ implies $n'_b = 0$. However, using only n_c or n'_d would be too uninformative since the converse of the implication does not hold. Table 1 shows that each of n_c, n'_d and n'_b capture different (though related) amounts of redundancy in the ILN and that each metric by itself fails to properly discriminate between the seven ILNs depicted in Figure 2. For example, n_c and n'_c treat a *Tree*, *Star* and *Line* as qualitatively equal but disagree on whether a *Full Mesh* is as good as a *Ring*. Consequently, to compute an overall estimated quality e_Q of an identity link network, we combine the three separate metrics by taking their av-

[4] The metric value indicates the negative impact of one or more missing links in an ILN.

Link-Network Quality Estimation				
ILN	Bridge	Diameter	Closure	Est. Quality
Ring	$B = 0 \quad n_b = 0.00$	$D = 3 \quad n_d = 0.56$	$C = 0.40 \;\; n_c = 0.60$	$\mathbf{e_Q} = \mathbf{0.61}$
Mesh	$B = 1 \quad n_b = 0.38$	$D = 3 \quad n_d = 0.56$	$C = 0.47 \;\; n_c = 0.53$	$\mathbf{e_Q} = \mathbf{0.51}$
Star	$B = 5 \quad n_b = 1.00$	$D = 2 \quad n_d = 0.38$	$C = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q} = \mathbf{0.32}$
Full Mesh	$B = 0 \quad n_b = 0.00$	$D = 3 \quad n_d = 0.00$	$C = 1.00 \;\; n_c = 0.00$	$\mathbf{e_Q} = \mathbf{1.00}$
Line	$B = 5 \quad n_b = 1.00$	$D = 1 \quad n_d = 1.00$	$C = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q} = \mathbf{0.11}$
Tree	$B = 5 \quad n_b = 1.00$	$D = 4 \quad n_d = 0.38$	$C = 0.33 \;\; n_c = 0.67$	$\mathbf{e_Q} = \mathbf{0.34}$

Table 1
Metrics values for each of the topologies from Fig. 2.

erage, 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 Intervals. The e_Q metric scores all ILNs on a continuous value in the [0,1] interval. To automatically discriminate potentially good networks from bad ones, we divide this interval into three segments: ILNs with values $0.9 \leqslant e_Q \leqslant 1$ will be rated as **good**, with values $0.75 < e_Q < 0.9$ as **undecided**, and with values $0 \leqslant e_Q \leqslant 0.75$ as **bad**. These boundaries are empirically determined, and can be adjusted depending on the use-case. The specific values of these boundaries does not affect the essence of our hypothesis.

Hypothesis. We can now state our hypothesis more formally: "*The* $\mathbf{e_Q}$ *intervals defined above are predictive of the quality of the links in an entity link network between multiple datasets*".

Example. By way of illustration, Table 1 gives the value of our e_Q metric for the six networks from Figure 2, and shows that the metric 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 in three experiments using between three to six datasets.

5. Datasets

We considered using datasets and gold standards from the OAEI initiative, but none of these go beyond links between two datasets. We therefore created our own gold standard on realistic datasets taken from the domain of social science, more specifically from the field of Science, Technology and Innovation studies. We consider this to be an important contribution of this paper. All datasets and our gold standard are available online at the locations given in later paragraphs.

Entities of interest to the STI domain of study are (among others) universities and other research-related organisations, such as R&D companies and funding agencies. Our six datasets are widely used in the field, and describe organisations and their properties such as name, location, type, size and other features.[5]

Grid[6] describes 80248 organisations across 221 countries using 12308 relationships. All organisations are assigned an address, while 96% of them have an organisation type, and only 78% have geographic coordinates.

OrgRef[7] collates data about the most important worldwide academic and research organisations (31000) from two main sources: Wikipedia and ISNI.

The Leiden Ranking dataset[8] 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[9] is a database on European Higher Education Institutions that not only includes research universities,

[5] 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. All these datasets are available on the RISIS platform at http://datasets.risis.eu/.

[6] https://www.grid.ac

[7] http://www.orgref.org

[8] http://www.leidenranking.com/

[9] https://www.eter-project.com/

but also colleges and a large number of specialized schools. The dataset covered 35 countries in 2015.

OrgReg[10] 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 36 countries.

The European Organisations' Projects H2020 database[11] documents the Horizon 2020 participating organisations.

6. e_Q Put to the Test

We test our hypothesis on a real life case study that revolves around the six datasets described in Section 5, 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 the other five datasets (Figure 3).

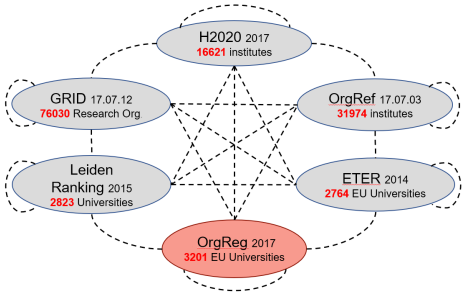


Fig. 3. Disambiguating OrgReg.
To evaluate $\mathbf{e_Q}$, all possible links are evaluated. So, the lack of one or more links is considered a potential evidence for suggesting the corresponding entities being different.

6.1. Experiment Design

Organizations are linked across or within datasets using an approximate string matching on their names with minimal similarity threshold 0.8. Based on this, we generate links between each pair of datasets, resulting in 21 sets of links (including linking a dataset to itself in order to detect duplicate entities in the dataset). We then take the union of all 21 sets of links, resulting in a collection of ILN's of varying size (see figure Figure 4).

Now that we have constructed a large collection of multi-dataset ILNs, we will compute the e_Q value for all of them. Then, the machine-predicted good/bad categories (using e_Q) will be checked against the ground truth by a non-domain expert (the first author of this paper) and further verified by a domain expert (the third author). This ground truth is available online.[12]

Notice that we have deliberately used a very weak entity resolution algorithm in this experiment (approximate string matching). This produces links of both very high and rather low quality, providing a genuine test for our e_Q metric to distinguish between them.

6.2. Results of first evaluation (non expert)

Ideally, we would find only ILNs of size 6 if each OrgReg entity were linked with one and only one entity in each of the five other datasets. With less than 100% coverage of OrgReg, we also expect to find ILNs of size smaller than 6. Figure 4 shows that we also find a substantial number of ILNs of size bigger than 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 (b) an imperfect matching algorithm (in our case approximate name matching), resulting in incorrect links in the ILN.

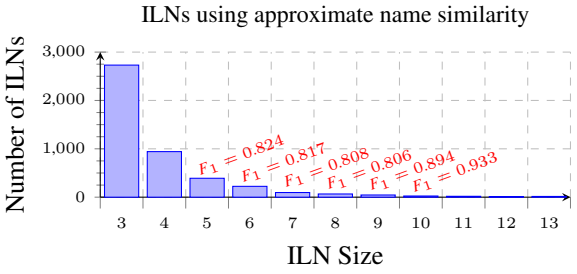


Fig. 4. Overview of the generated Identity Link Networks.

Due to the high number of ILNs generated[13], we evaluate only the 846 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 interval values for each of the 846 ILNs, and then compare the scores against those of a human

[10] http://risis.eu/orgreg/
[11] http://www.gaeu.com/sv/item/horizon-2020
[12] https://github.com/alkoudouss/Identity-Link-Network-Metric
[13] On a 6th Gen IntelÂő CoreâĎć i7 notebook with 8GB RAM, it takes about 1:40 minutes to automatically evaluate all 4398 clusters of size three and above (see Figure 4).

Majority Class Classifier (Baseline) vs Network Metric (e_Q) $\frac{MajorityClassClassifier}{NetworkMetrics}$ GT_P = Ground Truth Positive GT_N = Ground Truth Negative								
Size	$GT_P \mid GT_N$	F_1	ACC	NPV	$GT_P \mid GT_N$	F_1	ACC	NPV
3					56 | 8	**0.933** / 0.931	**0.875** / **0.875**	– / **0.5**
4					19 | 5	**0.884** / 0.878	**0.792** / **0.792**	– / **0.5**
5	272 | 119	0.821 / **0.824**	0.696 / **0.747**	– / **0.598**	14 | 1	**0.966** / 0.929	**0.933** / 0.867	– / **0**
6	139 | 85	0.766 / **0.817**	0.621 / **0.768**	– / **0.709**	14 | 5	**0.848** / **0.848**	**0.737** / **0.737**	– / –
7	50 | 56	0.685 / **0.808**	0.521 / **0.792**	– / **0.810**	10 | 2	0.909 / **1.0**	0.833 / **1.0**	– / **1.0**
8	35 | 31	0.693 / **0.806**	0.530 / **0.803**	– / **0.765**	4 | 0	**1.0** / **1.0**	**1.0** / **1.0**	– / –
9	21 | 24	– / **0.894**	0.533 / **0.889**	0.533 / **1**	8 | 1	0.941 / **1.0**	0.889 / **1.0**	– / **1.0**
10	8 | 16	– / **0.933**	0.667 / **0.958**	0.667 / **0.941**	1 | 0	**1.0** / **1.0**	**1.0** / **1.0**	– / –

Table 2

Network-metric (e_Q) results compared to the MCC baseline using non expert Ground Truth (left), and Expert sampled Ground Truth (right).

expert, resulting in F_1 scores. In red, Figure 4 displays the F_1 value for each ILN size. Overall, our e_Q metric resulted in high F_1 values ($0.806 \leqslant F_1 \leqslant 0.933$). We also pitched our e_Q metric against a Majority Class Classifier. Table 2 shows that our e_Q metric outperforms the Classifier on 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 (expert)

For a further evaluation by a Dutch domain expert from the field of STI (the third author of this paper), we selected 148 ILNs (ranging from size 3 to 10 as depicted in Table 2) in which at least one entity is located in the Netherlands. The expert deviated from the first evaluation in only 12 out of 148 cases. Although the changes slightly affect the ground truth for each ILN size, the F_1 values computed here are even higher ($0.848 \leqslant F_1 \leqslant 1$) as compared to the previous experiment. This shows that the non-expert nature of the first human judgement was not detrimental to our results. [14] This second experiment confirms our finding in the first experiment that e_Q is a reliable predictor of ILN quality.

[14] However, the very imbalanced character of the ground truth makes it hard to always outperform the baseline as illustrated in Table 2

6.4. Analysis

Both of the evaluations of e_Q above resulted in very high F_1 average values of 0.847 and 0.961 respectively. Furthermore, e_Q outperformed a majority-class classifier 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 quality of the links between the entities in an Identity Link Network.

7. e_Q Estimations in Noisy Settings

The previous experiment created links between entities using a rather weak entity resolution heuristic. This was an interesting setting because such weak matching 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 have been constructed using a more sophisticated matching heuristic, where we can control the amount of incorrect links in the ILNs. 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 contain such geo-coordinates for organisations. To test the performance of the e_Q metric at various levels of noise, we implement three sub-experiments where noise (the number

of false positive links) is introduced by decreasing the name similarity threshold from 0.8 (experiment 1) to 0.7 and by increasing the geographic proximity distance threshold as described in the next sub-section.

7.1. Experiment Design

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

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

Phase-2: Refine links. Each set of links is then refined by applying an approximate name comparison over the linked resources with a threshold of 0.7.

By now, we have **geo-only** (without name comparison) and **geo+names** sets of links, 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.

7.2. Strict vs. Liberal Clustering

To understand how link-networks are formed as we increase the geo-similarity distance, Figure 5 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_6 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.

These scenarios 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 less strict, two types of event may happen: (1) formation of new ILNs and/or (2) merging of ILNs. Table 3, shows that, in experiment 2, phenomenon (1) overtakes (2), which explains the increase in the number of ILNs as the near-by distance increases.

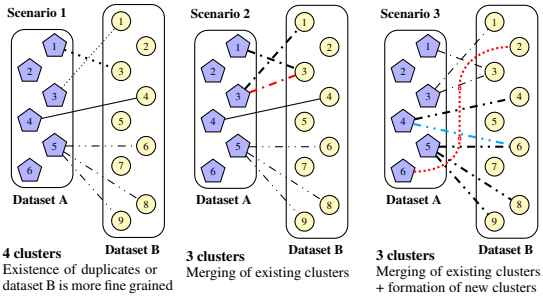


Fig. 5. Decrease/Increase of ILNs

7.3. Result and Analysis

Overall, as illustrated in Table 3, the number of ILNs generated in this experiment increases 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
$\geqslant 3$	230	36	738	168	841	371

Table 3
link-network overview.

Due to manpower limitations we restrict our evaluation efforts to networks of size 3. These ILNs cover 86% of the overall ILNs within 50m radius and 92% within 500m and 2k radius. Table 4 shows the results of pitching our e_Q metric against the human evaluation of the ILNs under both the geo-only and the geo+names conditions.

As an example, the values $F_1 = 0.803$ and $F_1 = 0.912$ depicted in the confusion matrices in Table 5 and Table 6 detail the machine quality judgements versus human evaluations of the networks generated within

	50 meters		**500 meters**		**2 kilometres**	
Size	geo-only	geo+names	geo-only	geo+names	geo-only	geo+names
$= 3$	92	31	249	155	198	342
Machine statistics on ILN's of size 3						
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						
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}:8 H_{bad}: 12
F_1 measures						
	$\mathbf{F_1 = 0.693}$	$\mathbf{F_1 = 0.826}$	$\mathbf{F_1 = 0.682}$	$\mathbf{F_1 = 0.909}$	$\mathbf{F_1 = 0.803}$	$\mathbf{F_1 = 0.912}$

Table 4

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 5

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 6

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

2 kilometres radius under respectively geo-only and geo+names conditions.[15]

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 proposed e_Q metric is in general able to exclude poor networks in noisy environments and to include good networks in noise-less environments. In addition, on the one hand, the relatively low F_1 measures displayed in Table 7 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 7 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

[15] All confusion matrices supporting the analysis can be found on the RISIS project website at http://sms.risis.eu/assets/pdf/metrics-link-network.pdf

Majority Class Classifier (Baseline) vs Network Metrics (e_Q) $\frac{MajorityClassClassifier}{NetworkMetrics}$ GT = Ground Truth GT_P = Ground Truth Positive GT_N = Ground Truth Negative					
50m geo-only	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 geo-only	GT=249	GT_P=66 GT_N=183	F_1 : $\frac{-}{\mathbf{0.682}}$	ACC: $\frac{0.735}{\mathbf{0.779}}$	NPV: $\frac{0.735}{\mathbf{0.978}}$
2km geo-only	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 geo+names	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 geo+names	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 geo+names	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 7
Network-metric (e_Q) result versus the MCC baseline.

takes the lead, the network metric does not fall far behind.

As 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 domain expert deviated from the first evaluation in only 1 case.

8. e_Q Put to a Ranking Test

The authors of the recently published paper [16] compared seven algorithms for clustering entities from multiple sources at different string similarity thresholds. They evaluated the quality of the clusters that these algorithms generated on three gold standard datasets[16], one manually built (referred here as GT1), and two syntactically generated. We take the evaluation results from [16] on GT1, and then test if our e_Q score is able to correctly predict the ranking of the algorithms as found in the reported evaluation. In contrast to the earlier experiments (where we use e_Q to assess the quality of clusters), we are now testing if e_Q can be used to correctly rank different clustering algorithms across datasets.

A slightly complicating factor is that the evaluation in [16] relies on F_1 values computed on *true pairs of entities found*. Since e_Q evaluates entire clusters (i.e. *sets* of pairs of entities) of size greater than 2 ($S > 2$), we recompute the F_1 values based on *true clusters found ($S > 2$)* and plot these performance measures for each algorithm in Figure 6 as *Baseline*. The resulting plot is comparable to the original one in [16]. We then ran the e_Q metric over the outputs of each algorithm at the same thresholds, displayed in Figure 6 as e_Q *Evaluation*.

[16] https://dbs.uni-leipzig.de/de/research/projects/object_matching/famer

The results show that the ranking of the algorithms by e_Q ($\mathbf{e_Q}$ ***Evaluation***) does not significantly deviate from the recomputed ranking of the algorithms as found in [16] (***Baseline***). To quantitatively support our findings, we have computed the F_1-based rankings error difference between the baseline and the four e_Q metrics and display it in Figure 7. Zoomed in, Figure 7 shows a deviation of ± 0.96 depending on the threshold (x axis) under which the clustering algorithms are evaluated. As illustrated, Figure 7 shows that, overall, the ranking error increases with the increase of the threshold, indicating that it becomes harder to discriminate between algorithms as the string similarity is set to tolerate less errors. Overall, the result illustrates the usefulness of the e_Q metric as it demonstrates its potential to rank (clearly dissociate) clustering algorithms whenever they show *significant performance differences*.

9. Refinements of e_Q Using Link Confidence Scores Produced by Entity Resolution Algorithms

Given that all links have been searched for, the absence of a link in an ILN network is shown to cripple the ideal structure of the network as it increases the chance for a *longer diameter* and the appearance of *bridges*, and it *reduces the density* of the network. These characteristics are thereby used by the e_Q metric as a potential evidence for tagging as GOOD or BAD the network as a whole. Furthermore, the metric assumes a ***link correctness confidence score of 1*** for all links in the network although it is not the case in the realm of entity matching unless a perfect match is found. Entity matching algorithms often produce pairwise matched entities with a confidence score in the interval [0, 1] as a quantitative justification for the pair to be the same.

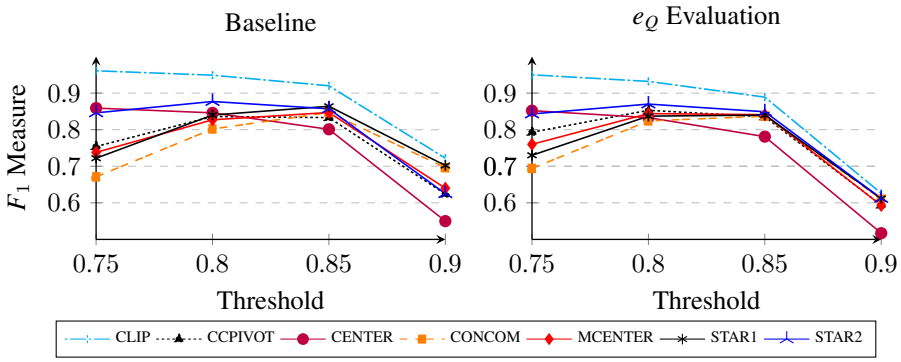


Fig. 6. Evaluation of e_Q on the ranking from [16]

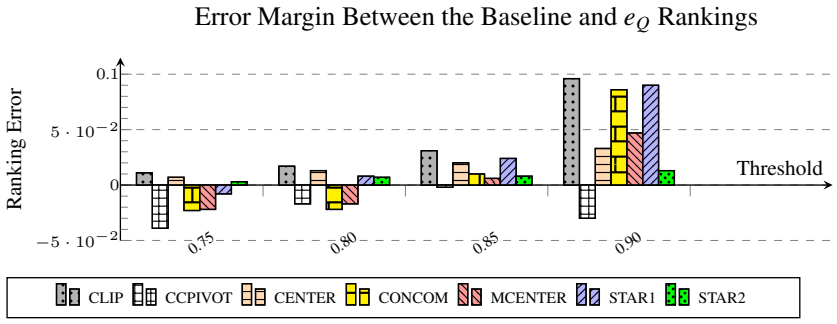


Fig. 7. Ranking deviations

So far, we strictly estimate the quality of an identity network based on the cost of its missing links and thereby its structure. Now, the wonder lies in *how to capture the toll of an existing link on estimating the quality of the network given that the link has a confidence score below one?* In other words, *is the strength of a identity link relevant in estimating the quality of the network using its structure?*

To understand the importance of the strength of links in estimating the quality of an identity network using its structure, we propose three new network quality estimation metrics ($e_{Q_{min}}$, $e_{Q_{avg}}$ and e_{Q_w}) that in their respective ways *combine structure and link strength* for network quality estimations. We evaluate these alternative metrics on the same ground truths used in sections 6 to 8, and compare each one of them to the original e_Q metric based on their respective F_1 scores in these various scenarios.

Before diving into the intricacies of link strength integration, we first start with the formalism that pave the way for understanding it.

A weighted, undirected, connected (WUC) graph[17] is defined as G = (V, L, w) where V is the set of nodes, L is the set of links or edges, and $w : L \mapsto \mathbb{R}^+$ is a function mapping edges $e_i = (v_{i-1}, v_i) \in L$ where $v_i \in V for\ i \in [1, k]$, unordered pair of vertices, to their decimal values $w(e_i)$ in the interval [0, 1]. The weight of sub-graph $H \subset G$ is $w(H) = \sum_{e \in L(H)} w(e)$ where $L(H)$ are the edges of H.

For two vertices a and $b \in V$, a path between a and b is a sequence $\pi = (e_1, e_2, ..., e_k)$ where $e_i = \{v_{i-1}, v_i\} \in L$ and $v_i \in V$ for $i \in \{1, ..., k\} = [1, k]$, with $v_o = a$ and $v_k = b$. $\Pi(a, b)$ denotes the set of all paths from a to b. The geodesic distance and weighted geodesic distance between a and b are respectively given by eqs. (1) and (2).

$$dist(a, b) = |\min(\pi \in \Pi(a, b))| \tag{1}$$

$$dist_w(a, b) = \min_{\pi \in \Pi(a,b)} \sum_{e \in \pi} w(e) \tag{2}$$

[17] We interchangeably refer to the undirected identity graph as network or cluster.

and the diameter and weighted diameter of G are given by eqs. (3) and (4)

$$diam(G) = \max\{dist(a,b), a,b \in V\} \tag{3}$$

$$diam_w(G) = \max\{dist_w(a,b), a,b \in V\} \tag{4}$$

We now have the prerequisites in place for presenting three hybrid ways of integrating link strength into the proposed network quality estimation metric.

9.1. Weakest Link

In this approach, we define $e_{Q_{min}}$ as the metric to estimate the quality of an identity network G based on both the structure of G and the strength of the links composing G. $e_{Q_{min}}$ is computed as *the product of the original e_Q score and the weakest link strength in the network* as given by eq. (5).

$$e_{Q_{min}} = e_Q \times \min_{e \in L(G)} w(e) \tag{5}$$

9.2. Link Average

Compared to the first weight integration approach, here, we simply replace the weakest link strength of G by the average of all strengths in G to obtain $e_{Q_{avg}}$ as provided in eq. (6).

$$e_{Q_{avg}} = e_Q \times \frac{\sum_{e \in L(G)} we(e_i)}{|L(G)|} \tag{6}$$

9.3. Rooted Link

As opposed to the first two approaches where we integrate the link strength without modifying the initial e_Q computation, here, we do the opposite. We use the link confidence score for computing each sub-metric score (bridge - diameter and closure). Doing so, the link confidence score is now more rooted into the initial e_Q formulation, leading to its equation adjustment. The detail on how the e_Q formula is adjusted for integrating the link's strength leading to Equation 10 is provided in the next paragraphs.

Weighted Bridge Metric. Given an identity graph G with n nodes, the idea here is to capture *the softening of the bridge metric measure as the strength of the edges composing the set of bridges in G weaken*. This is formulated in Equation 7: *the weaker the strength of a bridge gets, the less it negatively affects the quality of an identity network*.

$$n'_{b_w}(G) = max(n_{b_w}(G), sigmoid(w(B))) \tag{7}$$

where B is defined as sub-graph(s) of G whose edges are the bridges in G and $n_{b_w}(G) = \frac{w(B)}{n-1} = \frac{\sum_{e \in L(B)} w(e)}{n-1}$

Weighted Diameter Metric. Defined in Equation 8, the weighted diameter metric includes strength by *elongating the unweighted geodesic distance of G as the edges composing it weaken in strength*. In other words, *the smaller the strength, the longer the diameter gets*. This allows us to predict the decrease of the quality of an identity network whenever its diameter increases. It furthermore allows to increase the decrease of the quality of the identity network with respect to the weakening of the strength of each edge composing the network's diameter.

$$n'_{d_w}(G) = max(n_{d_w}(G), sigmoid(eDiam(G) - 1)) \tag{8}$$

where $eDiam(G) = 2diam(G) - diam_w(G) - 1$

and $n_{d_w}(G) = \begin{cases} 1 & \text{if } eDiam(G) > n-2 \\ \frac{eDiam(G)}{(n-1)-1} & \end{cases}$

For a better understanding of this measure, let assume two networks A and B with three edges. A with 3 nodes is a complete network and B with 4 nodes is a star network. For each network, the edges' strength are respectively $w(e_1) = 0.9$, $w(e_3) = 0.5$ and $w(e_3) = 0.3$. Regardless of the strength in each network, the unweighted diameter metric of A and B are respectively $n_d(A) = \frac{1-1}{3-2} = 0$ and $n_d(B) = \frac{2-1}{4-2} = 0.5$ while their weighted diameter metrics are $n_{d_w}(A) = \frac{2-0.3-1}{3-2} = \frac{0.7}{1} = 0.7$ and $n_{d_w}(B) = \frac{4-(0.3+0.5)-1}{4-2} = \frac{2.2}{2} = 1$. This example illustrates that the weaker the edges of a diameter, the longer the weighted diameter.

Weighted Closure Metric. is computed by normalizing of the sum of the weighted edges of G as provided in Equation 9.

$$n_{c_w}(G) = 1 - \frac{w(G)}{\frac{1}{2}n(n-1)} = 1 - \frac{\sum_{e \in L} w(e)}{\frac{1}{2}n(n-1)} \tag{9}$$

We are now able to compute a weighted e_Q as defined in Equation 10. Observe that each of the three metrics outputs a score in the interval [0,1]. Therefore, the overall measure is also in the interval [0,1].

$$e_{Q_w}(G) = 1 - \frac{n'_{b_w}(G) + n'_{d_w}(G) + n_{c_w}(G)}{3} \quad (10)$$

10. Weighted Metrics to the Test

Evaluation in noiseless settings. We now re-run the experiments conducted in section 6 using all metrics, namely e_Q, $e_{Q_{min}}$, $e_{Q_{avg}}$ and e_{Q_w} for estimating the quality of clusters of varying sizes: (a) size 5 to 10 for non expert ground truth and (b) size 3 to 10 for Dutch expert in Figure 8. The goal of these experiments is to find out which of the metrics performs well overall given the range of cluster sizes. This is done by comparing the metrics respective performances on the basis of their F_1 measures.

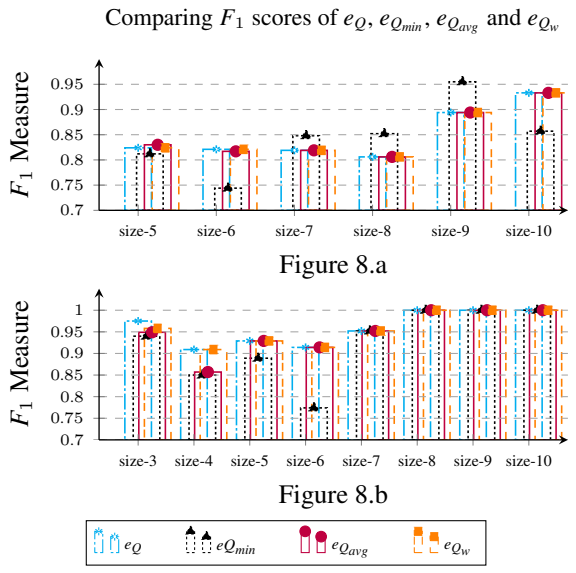


Fig. 8. Comparative evaluation in noiseless settings by a non expert (8.a) and Dutch domain expert (8.b)

Although in the expert evaluation, $e_{Q_{min}}$ and $e_{Q_{avg}}$ performed equally bad at least once, the observations show that, for both experiments, two main conclusions can be drawn: (1) $e_{Q_{min}}$ seams unreliable while (2) the rest of the metrics appear to perform alike, giving no solid indication on whether to combine structure and link confidence score.

Evaluation in noisy settings. Again here, we re-run the same experiments conducted in section 7 only now using all metrics. This, with the goal of comparing the metrics against each-other for further understanding the effect(s) of incorporating the link strength in the structure-based e_Q metric. Figure 9 again shows no solid evidence for being in favour of structure-based metric or "hybrid-based metrics" (structure + strength). The figure also shows that, whenever the identity network is composed of links with only confidence score of 1 (geo-similarity only), all approaches produce the same estimation score.

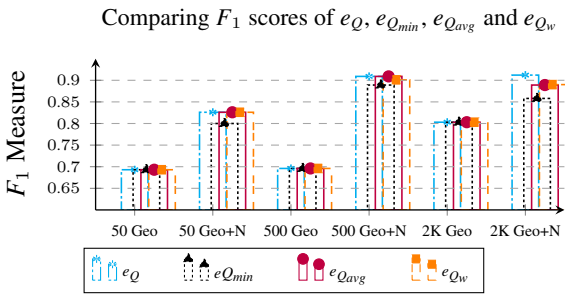


Fig. 9. Comparative evaluation by a domain expert in noisy settings

Evaluation for ranking clustering algorithms. Using data from [16], we show in Figure 10 the results of an experiment where we compare the ranking potential of each approach (e_Q, $e_{Q_{min}}$, $e_{Q_{avg}}$ and e_{Q_w}) for estimating the quality of an identity network against the algorithm rankings computed by Saeedi et al. (baseline). Bare in mind that here, we not only look at the performance in terms of F_1 measure but also in terms of ranking capability.

At first, the results show that all approaches appear to rank the algorithm almost equally. However, the deviation in terms of F_1 score for the $e_{Q_{min}}$ metric appears quite off compared to the baseline as it shifted considerably below the target's measures. With $e_{Q_{avg}}$, the previous F_1 measures move up but not yet close enough to those of the target. In the last option, which implements e_{Q_w} (Equation 10), the result is comparable to the target ranking and to the e_Q ranking as well, leading to a first judgement that the last two approaches perform better than the other two. According to the visualisation provided by Figure 10, e_Q and e_{Q_w} appear to be qualitatively comparable in performance with respect to the F_1 measures.

With a quantitative comparison provided by Table 8, the e_{Q_w} metric, followed by e_Q, appear to deviate from the baseline far less on average than the remaining ap-

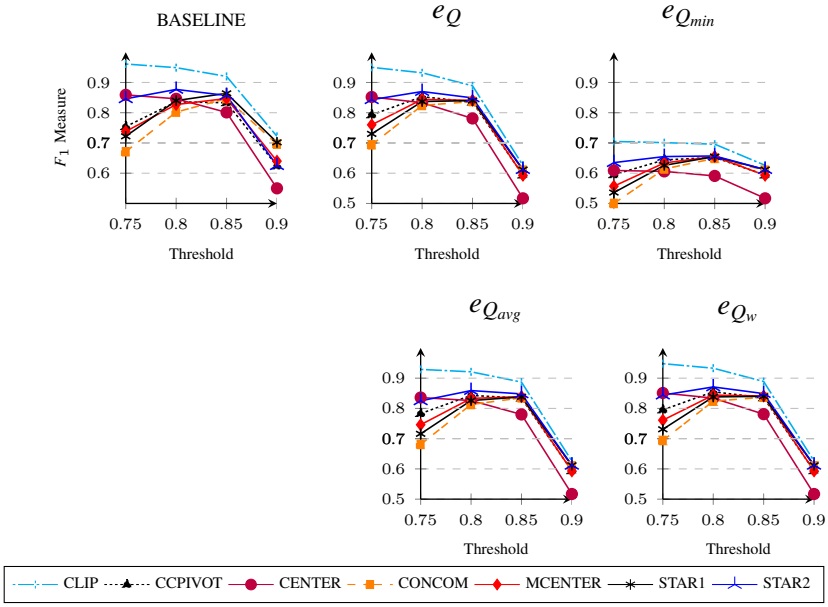


Fig. 10. Evaluation of e_Q on the ranking from [16]

proaches. This later observation shyly helps breaking the tie between the two metrics and indicates that e_{Q_w} is indeed the way to go as it provides an additional flavor that increases its information gain and thereby brings the measure closer to expressing *"how a network structure can be accurately translated into estimated quality"*.

Discussion. Although our last experiment seams in favour of the e_{Q_w} metric, truth is, we need more experiments for making a convincing case on whether one of the hybrids methods is worth the extra computation compared to the original metric, or whether a specific hybrid method works best in some particular settings. For example, we suspect that in settings where matching algorithms are rather permissive, there should be compelling reasons for the link strength to be included. Perhaps, in this situation where link confidence could be assigned a score in the range [0.3, 1] for example, even $e_{Q_{min}}$ could turn up stable. This, because, in our scenarios, we filter the potentially good links prior to estimating the quality of the network they form. Now, what if this task is given to the metric?

11. Conclusion and Future Work

11.1. Conclusion

Entity resolution is an essential step in the use of multiple datasets on the semantic web. Since entity resolution algorithms are far from being perfect, the links they discover must often be human validated. Because 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 datasets, using a combination of graph metrics over the network (> 2) formed by these links. Our metric captures the intuition that high redundancy in such a linking-network correlates with high quality. Furthermore, we have proposed hybrid-metrics that combine structure and link confidence score for the same purpose of estimating the quality of links between entities. The intuition here is an incremental improvement of the original metric by evaluating the integration or not of link strength in its estimation computation.

We have tested our metric in three different scenarios. Using a collection of six widely used social science datasets in the first two experimental settings, we compared the predictions of link quality by our metric against human judgements on hundreds of networks involving thousands of links. In both evaluations, our metric correlated strongly with human judgement ($0.806 \leqslant F_1 \leqslant 1$), and it consistently beats the Majority Class Classifier baseline (except in cases where this is numerically near impossible because of a highly skewed class distribution). In the experimental

THRESHOLD	0.75	0.80	0 .85	0.90	AVERAGE
CLIP BASELINE	0.961	0.949	0.920	0.722	0.888
e_Q	0.011	0.017	0.031	0.096	**0.039**
$e_{Q_{min}}$	0.256	0.248	0.224	0.096	0.206
$e_{Q_{avg}}$	0.032	0.028	0.033	0.096	0.047
e_{Q_w}	0.013	0.016	0.031	0.096	**0.039**
CCPIVOT BASELINE	0.754	0.836	0.833	0.624	0.762
e_Q	-0.039	-0.017	-0.002	0.03	**-0.007**
$e_{Q_{min}}$	0.159	0.192	0.181	0.03	0.141
$e_{Q_{avg}}$	-0.029	-0.007	-0.001	0.03	**-0.002**
e_{Q_w}	-0.041	-0.019	-0.002	0.03	-0.008
CENTER BASELINE	0.859	0.846	0.801	0.55	0.765
e_Q	0.007	0.016	0.02	0.033	**0.019**
$e_{Q_{min}}$	0.25	0.24	0.21	0.033	0.183
$e_{Q_{avg}}$	0.023	0.02	0.021	0.033	0.024
e_{Q_w}	0.008	0.012	0.02	0.033	**0.018**
CONCOM BASELINE	0.671	0.802	0.847	0.696	0.754
e_Q	-0.023	-0.022	0.01	0.086	**0.013**
$e_{Q_{min}}$	0.171	0.188	0.198	0.086	0.161
$e_{Q_{avg}}$	-0.01	-0.011	0.012	0.086	0.019
e_{Q_w}	-0.023	-0.022	0.01	0.086	**0.013**
MCENTER BASELINE	0.738	0.827	0.847	0.64	0.763
e_Q	-0.022	-0.017	0.006	0.047	**0.004**
$e_{Q_{min}}$	0.181	0.193	0.192	0.047	0.153
$e_{Q_{avg}}$	-0.008	-0.023	-0.053	0.047	-0.009
e_{Q_w}	-0.023	-0.017	0.006	0.047	**0.003**
STAR1 BASELINE	0.722	0.84	0.864	0.702	0.782
e_Q	-0.008	0.003	0.024	0.09	**0.027**
$e_{Q_{min}}$	0.186	0.214	0.209	0.09	0.175
$e_{Q_{avg}}$	0.006	0.014	0.026	0.09	0.034
e_{Q_w}	-0.009	0.002	0.024	0.09	**0.027**
STAR2 BASELINE	0.846	0.877	0.857	0.625	0.801
e_Q	0.003	0.007	0.008	0.013	**0.008**
$e_{Q_{min}}$	0.211	0.222	0.2	0.013	0.162
$e_{Q_{avg}}$	0.021	0.018	0.009	0.013	0.015
e_{Q_w}	0.002	0.006	0.008	0.013	**0.007**

Table 8

Comparing the ranking capability of each of the e_Q approaches. For each algorithm, we compare the baseline F_1 scores to those of an e_Q approach, and only report the difference. Then, for each approach, we compute by how much the e_Q metric scores under scrutiny deviate on average from those of the baseline. Using the later average, we compare the e_Q approaches against each other.

condition where we deliberately constructed noisy and non-noisy link-networks, we showed that our metric is in general able to exclude poor networks in noisy environments and to include good networks in noiseless environments. With the last experiment, we also show that our metric is able to rank entity resolution algorithms on their quality, using an externally produced dataset and corresponding ground truth. All this amounts to testing the e_Q metric on a dozen different algorithms and parameter settings.

After showing that, across these different experimental conditions, our quality metric consistently agrees with human judgement, we re-run all experiments on both the e_Q and hybrid metrics. The results suggest that the hybrid methods seam to have an effect on estimating the quality of an identity network, only it is yet unclear in what specific condition(s) these metrics bear fruit (do significantly well as opposed to e_Q). This yells for more experiments on the matter.

To encourage replication studies and extensions to our work, all the datasets used in these experiments are available online.

11.2. Future work

Networks of size two. The presented metrics are shown to work well in clusters of size bigger than two. Finding ways in which networks of size two can be validated using the e_Q metrics would be an added value as the amount of clusters of such size is not negligible.

Dynamic link adjustment. The current work ideally takes clustered ILNs networks as input. However, when such networks are not provided, it simply takes the output of an entity resolution algorithm as given, applies a simple clustering algorithm (section 2) and tries to estimate the quality of that output. A closer coupling between our metric and an entity resolution algorithm would allow the algorithm to dynamically adjust its output based on the e_Q quality estimates. Similarly, embedded in a user-interface, the score of our metric could help the user to give the final judgement to accept or reject an ILN.

Parameter tuning. In this work, we empirically determined the 1.6 sigmoid hyper-parameter, the discrete e_Q intervals and the string similarity thresholds. Experimenting on fine-tuning these parameters using the current ground truths and data from other domains would help understanding how and when different choices could lead to an increase or a decrease of the metrics' predictive power.

11.3. Acknowledgement

We kindly thank *Paul Groth* for his constructive comments and proofreading, *Alieh Saeedi* for sharing her experiments data and supporting the reproducibility of their experiments, and the EKAW *reviewers* for constructive comments. This work was supported by the European Union's 7th Framework Programme under the project RISIS (GA no. 313082).

References

[1] A.K. Idrissou, F. van Harmelen and P. van den Besselaar, Network Metrics for Assessing the Quality of Entity Resolution Between Multiple Datasets, in: *Knowledge Engineering and Knowledge Management*, C. Faron Zucker, C. Ghidini, A. Napoli and Y. Toussaint, eds, Springer International Publishing, Cham, 2018, pp. 147–162.

[2] 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*, Springer, 2014, pp. 457–471.

[3] A.-C.N. Ngomo and S. Auer, Limes-a time-efficient approach for large-scale link discovery on the web of data., in: *IJCAI*, 2011, pp. 2312–2317.

[4] O. Hassanzadeh, R. Xin, R.J. Miller, A. Kementsietsidis, L. Lim and M. Wang, Linkage query writer, *Proceedings of the VLDB Endowment* **2**(2) (2009), 1590–1593.

[5] 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*, ACM, 2009, pp. 1027–1036.

[6] J. Volz, C. Bizer, M. Gaedke and G. Kobilarov, Discovering and maintaining links on the web of data, in: *ISWC*, Springer, 2009, pp. 650–665.

[7] J. Euzenat and P. Shvaiko, *Ontology matching*, 2nd edn, Springer-Verlag, Heidelberg (DE), 2013.

[8] A. Maedche and S. Staab, Measuring similarity between ontologies, in: *International Conference on Knowledge Engineering and Knowledge Management*, Springer, 2002, pp. 251–263.

[9] D. Vrandečić and Y. Sure, How to design better ontology metrics, in: *ESWC*, Springer, 2007, pp. 311–325.

[10] J. David and J. Euzenat, Comparison between ontology distances (preliminary results), in: *ISWC*, Springer, 2008, pp. 245–260.

[11] J. David, J. Euzenat and O. Šváb-Zamazal, Ontology similarity in the alignment space, in: *ISWC*, Springer, 2010, pp. 129–144.

[12] W. Li, S. Zhang and G. Qi, A graph-based approach for resolving incoherent ontology mappings, in: *Web Intelligence*, Vol. 16, IOS Press, 2018, pp. 15–35.

[13] C. Sarasua, S. Staab and M. Thimm, Methods for intrinsic evaluation of links in the web of data, in: *ESWC*, Springer, 2017, pp. 68–84.

[14] A. Baron and M. Freedman, Who is who and what is what: experiments in cross-document co-reference, in: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, 2008, pp. 274–283.

[15] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, in: *Proceedings of the 2007 Joint Conference on EMNLP-CoNLL*, 2007.

[16] A. Saeedi, E. Peukert and E. Rahm, Using Link Features for Entity Clustering in Knowledge Graphs, in: *ESWC*, Springer, 2018, pp. 576–592.

[17] C. Guéret, P. Groth, C. Stadler and J. Lehmann, Assessing linked data mappings using network measures, in: *ESWC*, Springer, 2012, pp. 87–102.

[18] D. Menestrina, S.E. Whang and H. Garcia-Molina, Evaluating entity resolution results, *Proceedings of the VLDB Endowment* **3**(1–2) (2010), 208–219.

[19] S. Cucerzan, Large-scale named entity disambiguation based on Wikipedia data, in: *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[20] A.K. Idrissou, R. Hoekstra, F. van Harmelen, A. Khalili and P. van den Besselaar, Is My:SameAs the Same As Your:SameAs?: Lenticular Lenses for Context-Specific Identity, in: *Proceedings of the Knowledge Capture Conference*, K-CAP 2017, ACM, New York, NY, USA, 2017, pp. 23:1–23:8. ISBN 978-1-4503-5553-7. doi:10.1145/3148011.3148029.

[21] E. Ruckhaus, O. Baldizán and M.-E. Vidal, Analyzing linked data quality with liquate, in: *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, Springer, 2013, pp. 629–638.