Applied Network Science

RESEARCH Open Access

Network-theoretic information extraction quality assessment in the human trafficking domain

Mayank Kejriwal* and Rahul Kapoor

*Correspondence: kejriwal@isi.edu
Information Sciences Institute,
University of Southern California,
4676 Admiralty Way, Ste. 1001,
Marina del Rey, California, USA

Abstract

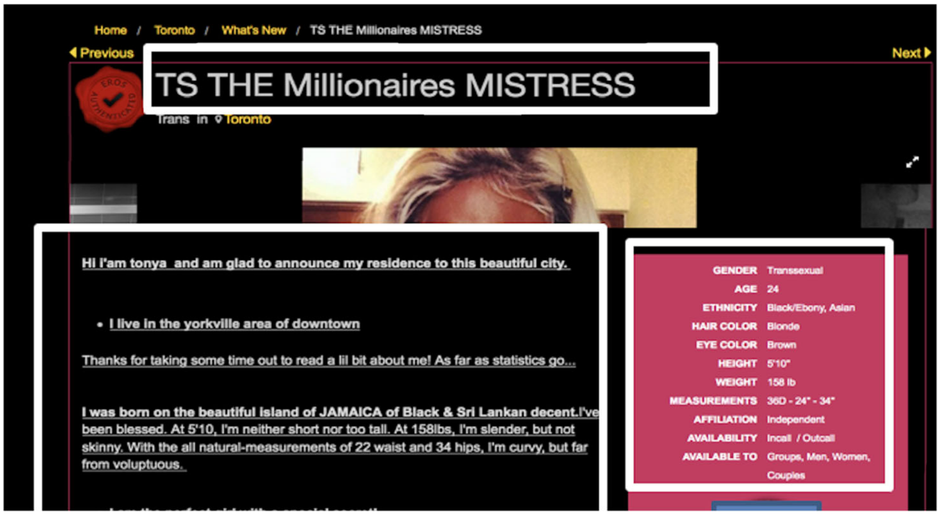
Information extraction (IE) is an important problem in Natural Language Processing (NLP) and Web Mining communities. Recently, IE has been applied to online sex advertisements with the goal of powering search and analytics systems that can help law enforcement investigate human trafficking (HT). Extracting key attributes such as names, phone numbers and addresses from online sex ads is extremely challenging, since such webpages contain boilerplate, obfuscation, and extraneous text in unusual language models. Assessing the quality of an IE system is an important problem that is particularly problematic in this domain due to lack of gold standard datasets. Furthermore, building a robust ground truth from scratch is an expensive and time-consuming task for social scientists and law enforcement to undertake. In this article, we undertake the empirical challenge of analyzing the quality of IE outputs in the HT domain without the provision of laboriously annotated ground truths. Specifically, we use concepts from network science to construct and study an extraction graph from IE outputs collected over a corpus of online sex ads. Our studies show that network metrics, which require no labeled ground truths, share interesting and consistent correlations with IE accuracy metrics (e.g., precision and recall) that do require ground-truths. Our methods can potentially be applied for comparing the quality of different IE systems in the HT domain without access to ground-truths.

Keywords: Information extraction, Structural analysis, Human trafficking, Relational analysis, Network theory, Attributed networks, Artificial intelligence

Introduction

Information extraction (IE) is a broad area in both the Natural Language Processing (NLP) and the Web communities (Chang et al. 2006a, b). The main goal of IE is to extract useful information from raw documents and webpages. Traditional IE, which is assumed in this article, assumes a particular schema according to which information must be extracted and typed. Domain-specific applications, such as human trafficking, generally require the schema to be specific and fine-grained, supporting attributes of interest to investigators, including phone number, address and also physical features such as hair color and eye color (Fig. 1). As shown in the figure, some attributes may occur as 'links' (e.g., phone number) and are not directly visible in the text on the page. There is also considerable heterogeneity, both across webpages in the same Web domain (e.g., two individual webpages from backpage.com), and across Web domains (e.g., backpage.com

Springer Open



TS THE Millionaires MISTRESS

Trans in Toronto

Hi I'am tonya and am glad to announce my residence to this beautiful city.

- I live in the yorkville area of downtown

Thanks for taking some time out to read a lil bit about me! As far as statistics go...

I was born on the beautiful island of JAMAICA of Black & Sri Lankan decent.I've been blessed. At 5'10, I'm neither short nor too tall. At 158lbs, I'm slender, but not skinny. With the all natural-measurements of 22 waist and 34 hips, I'm curvy, but far from voluptuous.

GENDER	Transsexual
AGE	24
ETHNICITY	Black/Ebony, Asian
HAIR COLOR	Blonde
EYE COLOR	Brown
HEIGHT	5'10"
WEIGHT	158 lb
MEASUREMENTS	36D - 24" - 34"
AFFILIATION	Independent
AVAILABILITY	Incall / Outcall
AVAILABLE TO	Groups, Men, Women, Couples

Fig. 1 A truncated example of a webpage (images/key information are obfuscated) from the online sex advertisement domain. White boxes indicate regions of the webpage containing critical attributes (e.g., age) that need to be processed and extracted by an information extraction (IE) system

and liveescortreviews.com). All of these observations make IE a challenging problem in an illicit Web domain such as online sex trafficking.

As with other AI approaches, quality tradeoffs of rival IE systems are determined by applying them (after the appropriate training and validation, if applicable[1]) to a withheld (but still manually labeled) test dataset (Freitag 2000). It is less clear how IE quality can be evaluated without access to such ground-truths. This article probes the issue of whether it is possible to use the relational structure of an IE system's outputs to characterize its quality (using well defined metrics) *without* the provision of ground-truths. Rather than focus on a theoretical model, the article specifically considers quality evaluation in the online sex advertisement domain. We hypothesize that, as with other relational systems, *network science* could be used in support of this goal.

We exploit the following intuition in support of this goal. Since it is generally the case that an attribute (such as city) is not extracted from a single document, but from multiple documents, extractions tend to be 'shared' between documents. Furthermore, a single document can yield more than one extraction per attribute, especially if the underlying IE system is recall-favoring, and some extraction combinations are higher-probability than others. For example, 'Charlotte' and 'Raleigh' have higher probability of being extracted from the same document than 'Charlotte' and 'Los Angeles'. In the same vein, some extractions are more noise-prone than others e.g., Charlotte has higher potential to be mis-extracted as a name (in some documents) than Raleigh. Such relational connections can be used to model the set of documents and IE extractions as an *attribute extraction network*. The AEN is constructed by modeling extractions as nodes in the network, and by modeling shared extractions (within a single Web document) as network edges. Our overarching hypothesis is that changes in the structure of this simple network can be used to 'track' changes in the quality of the underlying IE system that yielded the extractions in the first place. Specifically, we propose to answer the following research questions for a domain-specific corpus of online sex advertisements:

Research Question (RQ) 1 : When extractions from a corpus are represented as an attribute extraction network (AEN), how does the structure of the network change as the quality of the underlying IE system changes, where structure is measured using single-point network-theoretic metrics such as algebraic connectivity and network diameter?

Research Question (RQ) 2: How does the degree distribution of an AEN change as IE quality changes? Are degree distributions normal?

We note that, strictly speaking, RQ2 is a special case of RQ1, since the degree distribution is a function of network structure as well. However, for methodological reasons we choose to separate the two questions with the first question covering single-point metrics (such as connectivity metrics, diameter etc.) and the second, covering the degree distribution itself. Potentially, any distribution (e.g., clustering coefficient distribution) could be selected for investigation in lieu of the degree distribution. However, the motivation in choosing the degree distribution is to specifically investigate if networks such as the AEN obey (approximately) normal distributions. If not, then important concerns arise as to whether simple random sampling and labeling is appropriate when constructing a gold standard (or training dataset) for an IE. In fact, many theoretical treatments on machine learning make critical assumptions about i.i.d (independent and identically distributed) data. Similarly, when evaluating machine learning and NLP systems, it is often the case that (when exhaustive ground-truths are not available) outputs are randomly sampled and annotated, the hope being that the measured performance will generalize statistically with finite, but sufficiently sized, samples.

By plotting the degree distribution of the AEN, we can directly analyze whether noise in the IE system is i.i.d by verifying that the degree distribution is Gaussian. If, instead, the distribution exhibits a power-law trend as with scale-free networks, for example, non-normality would be strongly indicated, implying that we should revise our statistical assumptions when deciding how to sample and annotate extractions for IE quality assessments. Furthermore, by studying how (or whether) the degree distribution changes or undergoes drift as the level of noise in the IE outputs increases, we hope to gain interesting and direct insights into the nature of IE noise.

For conducting these empirical studies, we use sex advertisement data scraped from the Open Web, and attributes extracted by a relatively advanced IE for an in-use investigative search engine developed in our previous work (Kejriwal and Szekely 2017b). This search engine is being used by multiple investigative agencies in the United States, and empirical work conducted in support of this article is being directly applied (Kejriwal et al. 2018). However, while our previous work was focused on describing and evaluating a search engine for HT, as well as the information extraction programs that fed into the search engine, this work centers on evaluating IE on an HT-specific corpus without access to a ground truth.

Contributions

Specific contributions in this article are follows. We propose to empirically study Information Extraction (IE) quality in the human trafficking (HT) domain using a novel network science-based framework, without relying on traditionally required ground truths. This HT-focused empirical study is a central contribution of the article, since an important motivation for our research is to mitigate the expense of acquiring a laboriously annotated ground-truth, without which an IE cannot be evaluated (and consequently, any system that relies on IE cannot be used with confidence by investigators). Specifically, we empirically study two research questions, using a 10,000+ document corpus of sex advertisements crawled from backpage.com[2], and a variety of IE systems executed over three attributes (name, city and phone). Our results show that there is a definite and consistent correlation across standard quality metrics as defined by the IE community, and structural metrics defined by the network science community. To the best of our knowledge, such a correlation has never been noted or exploited before in prior work. Our results also suggest the possibility of using structural metrics, which can be deduced in an unsupervised manner without access to a ground truth, to study whether a given IE is deviating from the ground truth (compared to another IE) on a quality metric such as precision or F-measure.

Related work

The primary problem that is being studied in this article is the evaluation of an important AI approach (information extraction) in a domain with deep social impact (human trafficking). Because we cannot assume a ground-truth, the evaluation is conducted using network-theoretic techniques. All of these individual fields of study (information extraction, computational investigations of human trafficking , and network science) have individually received much research attention, as we describe in the sub-sections below. However, we are not aware of any network science papers that are related to evaluation of IE quality without potential access to ground truths, either within or without the context of human trafficking. This article attempts to build such a bridge within the context of the special domain of human trafficking.

Information extraction

Information extraction is a core component of any information integration pipeline over Web and natural language corpora, as 'unstructured'[3] data must first be rendered into a machine-readable, structured form in order for fine-grained queries to be executed over them. With the initial advent of the Web, wrapper induction systems had proved successful for several IE domains (Kushmerick et al. 1997). State-of-the-art work in the early 2000s (e.g. STALKER (Muslea et al. 1998)) used machine learning methods for the wrapper induction problem (Lerman et al. 2003). Such methods were inherently data-driven, and were less brittle than rule-based wrapper architectures. IE systems have continued to evolve since then; Chang et al. provide a comparative survey of many of the leading IE techniques along three dimensions (task domain, degree of automation and the actual techniques used) (Chang et al. 2006a). A key finding of the survey is the dependence of techniques on the actual input format. For example, while unsupervised and semi-supervised methods are well-suited for template pages, regular expressions and supervised approaches tend to be more robust for non-template pages (Lerman et al. 2003; Muslea et al. 1998). A consequent problem arising from such diverse methodologies is evaluating precision and recall in a consistent way (Chang et al. 2006a).

There has been much research on IE in traditional domains, and on datasets that are 'well-formed' (e.g., newswire) with accuracy on attributes such as person names often exceeding 80% (Nadeau and Sekine 2007). In contrast, it is well known that for more complex extractions (including relation and event extractions), accuracy is much lower (Ahn 2006). A similar problem occurs when one moves from newswire to social media and unusual domains that have not been well-studied, either socially or computationally (Ritter et al. 2012).

Evaluation of IE

As with IE systems development, IE evaluation (in the research community) was also predominantly designed for newswire-resembling corpora, with competitions and efforts such as the Message Understanding Conference (MUC) and Text Retrieval Conference (TREC) series involving the annotation of large corpora of data to ensure sufficient resources for training, validation and testing (Chinchor 1998; Voorhees and et al 1999). For illicit domains such as human trafficking, ground-truths do not exist and are hard to acquire. We note also that one cannot crowdsource the annotations due to the sensitivity of sex advertisement data. The cost of labeling is also an issue, since investigative agencies are typically resource-strapped to begin with (and cannot dedicate additional resources to an annotation service).

Domain-specific applications of IE

IE has many applications, one of which is knowledge graph construction (KGC). KGC draws on advances from a number of different research areas, including information extraction (Chang et al. 2006b), information integration (Doan et al. 2012), and inferential tasks such as entity resolution (Elmagarmid et al. 2007). Good examples of *architectures* that implement KGC principles are Domain-specific Insight Graphs (DIG) and DeepDive (Niu et al. 2012; Szekely et al. 2015). Both of these architectures have a significant IE component, and also rely (either directly or indirectly) on the quality of the extractions in important sub-components such as search and analytics.

More recently, Open Information Extraction or OpenIE has become a popular topic of research, owing to the need for IE techniques that do not rely on pre-specified vocabularies (Banko et al. 2007; Etzioni et al. 2008). In a preliminary version of the system, we tried state-of-the-art versions of OpenIE, including both old and new versions of the system proposed by (Etzioni et al. 2008). Even when relevant extractions were obtained from the corpus of webpages, the precision and recall were both judged to be too low to be useful. This largely motivated our earlier research on focused knowledge graph construction for illicit domains, albeit only for keyword queries that were easily amenable to GUI integration (Szekely et al. 2015).

The importance of domain-specific IE has also been rising through a series of ambitious projects. For example, the Defense Advanced Research Projects Agency (DARPA) MEMEX program[4], which funded multiple institutions in the United States to build semi-automatic, democratized domain-specific search systems, led to national efforts in using such technology for combating human trafficking[5]. IE was a crucial step in setting up such search engines.

Beyond human trafficking, investigators in other illicit domains, such as narcotics, securities fraud and illegal weapons sales, also expressed interest in using the technology. However, before one can deploy IE and search technology to such agencies, it is important to get some sense of the quality of multiple IE systems, and to also reason about changes in quality with the tuning of parameters. For example, when extracting attributes from text that has been scraped from webpages, it is intuitively plausible (and empirically the case as well (Kapoor et al. 2017)), that the more text is scraped from the webpage, the higher will be the recall of an extraction system's output compared to the output if it were run on more conservatively parsed text. Precision tends to suffer, however, since extraneous text gets scraped and causes noise to creep into downstream extractors (such as a phone number extractor executed on the scraped text).

An empirical study on capturing such tradeoffs systematically, particularly without access to ground-truths, has thus far been lacking. This article attempts to address this need. Specifically, in contrast with much of the prior work on IE, this work neither proposes a new system nor algorithm, but instead describes a network science-based framework that allows the evaluation and comparison of IE systems (for the HT domain) without being restricted by the availability of large quantities of labeled data. Furthermore, the empirical data and findings described in this article shed new insights on the nature of IE noise e.g., our evidence suggests that the i.i.d (independent and identically distributed) assumption often used in machine learning may not be applicable to IE in the HT domain.

Human trafficking (HT)

One of the most important aspects that separate this work from prior work is its focus on a non-traditional domain such as human trafficking (HT) that has an outsize presence on the Web. By some estimates, HT is a multi-billion dollar industry; however, due to both technical and social reasons, it has largely been ignored by the computational sciences till quite recently (Alvari et al. 2016; Hultgren et al. 2016). A notable exception in the knowledge graph construction domain is the DIG (Domain-specific Insight Graphs) system (Szekely et al. 2015). Similar to other systems such as DeepDive, DIG implements KGC components, in addition to a GUI, and was evaluated on human trafficking data. Those evaluations largely motivated this article, since extensive effort had to be expended to annotate even small ground truths.

More broadly, semi-supervised and minimally supervised AI has been applied to fight human trafficking in contexts beyond information extraction and search (Alvari et al. 2017; Burbano and Hernandez-Alvarez 2017; Kejriwal et al. 2017; Rabbany et al. 2018). As one example, the FlagIt system, recently developed in our group, attempts to semi-automatically mine indicators of human trafficking (which include movement, advertisement of multiple girls etc.) (Kejriwal et al. 2017). As another example, Rabbany et al. (2018) explore methods for active search of connections in order to build cases and combat human trafficking. Finally, although this work deals primarily with linguistic data (since it is focused on IE, which tends to work on linguistic data), there has also been a steady stream of work on the non-linguistic characteristics of sex ads. For example, recently, Whitney et al. describe how emojis can be used to add a layer of obfuscation to sex ads to avoid getting investigated, caught and prosecuted (Whitney et al. 2018). In part, this work is motivated by such findings: even if investigators invested the effort to painstakingly construct ground-truths, the creative and dynamic ways in which traffickers adapt (e.g., by using obfuscations such as emojis and misspellings) would soon render those ground-truths stale and obsolete. Hence, there is a real need for developing end-to-end unsupervised IE systems, both for acquiring and evaluating extractions.

Furthermore, although the research described herein is specifically designed to investigate and combat human trafficking, we believe that the core elements of the overall problem and solution can be extended to other domains (e.g. from the Dark Web Chen (2011)) that are highly heterogeneous, dynamic and that deliberately obfuscate key information. Because illicit domains are under-studied, and obtaining both raw and ground-truth data are difficult, we use a rich trove of documents available to us from the human trafficking domain to study the research problems in this article. Currently, it is too early to tell if the findings can be empirically extended to other illicit domains. On the other hand, multiple illicit domains share both common challenges (e.g., information obfuscation), and common needs (e.g., prioritization on extracting location-specific and identifier-specific attributes to assist law enforcement). These commonalities suggest that some of our empirical findings may be generalizable to other potential illicit domains such as securities fraud and narcotics.

Finally, in the context of this paper, it is important to distinguish between causation, correlation and prediction. Many of the results we explicitly describe are correlates; our research question, in fact, can be framed in terms of finding metrics that do not require labeled data but that are correlated with actual performance (captured properly by metrics that do require labeled data as a gold standard). However, we invoke a longitudinal argument in claiming that, because the networks are constructed from extractions, they are derivatives of real data and cannot (arguably) have caused the relational dependencies between extractions. We do not claim causation in any form, however, only that the network metrics are predictive of accuracy metrics by virtue of the correlation. More formal models that go into depth into such theoretical issues were presented in (Hultgren et al. 2016, 2018; Whitney et al. 2018).

Network science

Network science is an actively researched, standard framework for studying complex systems that possess structure (Barabási and et al 2016). Such systems include networks of protein-protein interactions (Gavin et al. 2002), citation networks (Hummon and Dereian 1989) and social networks (Borgatti et al. 2009), to only name a few. Recent research has led to many exciting advances in the construction and study of complex networks, especially from 'Big Data'. For example, Chen and Redner study the community structure of the physical review citation network from the mid-1890s to 2007 (Chen and Redner 2010). Other domain-specific examples include the study of patent citation networks in nanotechnology (Li et al. 2007) and the creation and influence of citation distortions (Greenberg 2009).

Another highly active sub-area of research in network science, and (arguably) one of the original motivations for employing network science as a scientific methodology for studying structure, is social networks. Work in this area can be traced back to at least the 1940s (and possibly beyond), when Moreno first proposed the 'sociogram' as a way of studying such systems at a structural level (Moreno 1946). Since then, there have been tens of thousands of papers and articles on the subject; a standard, highly comprehensive treatment on social network analysis was provided by Wasserman and Faust (Wasserman and Faust 1994), with a more recent book by Knoke and Yang (2008). More recently, pioneering work in this area include a study of networks, crowds and markets by Easley and Kleinberg (Easley et al. 2010), social tie inference in heterogeneous networks (Tang et al. 2012), prediction of positive and negative links in social networks (Leskovec et al. 2010) and even ethics and privacy-related challenges in mining social network data (Kleinberg 2007). Other important applications of network science include bioinformatics, with research ranging from studies in systems pharmacology (Berger and Iyengar 2009) to tools designed for fast network motif detection (Wernicke and Rasche 2006), (Schreiber and Schwöbbermeyer 2005).

This article is differentiated from the papers above by not attempting to use network science to study the properties of a domain by modeling its structure as a network; rather, we hypothesize that network science can be used to deduce (at least as a correlate) the levels of noise and data quality in real-world IE systems applied to consequential domains such as human trafficking. In that sense, the article presents a novel application of network science compared to prior related work.

Technical preliminaries

In this section, we introduce the necessary technical preliminaries to place the (subsequently described) empirical studies in context. Because the formalism is interdisciplinary and relies on both IE and network science, two constructs that do not traditionally intersect in the academic literature, we define concepts from the ground-up.

The core elements in our framework are documents, which in our specific application are blocks of text scraped from sex advertisements. A raw document D may be considered to be a pair $(id, text)$, where id is an identifier for the document, and $text$ is usually just a (potentially long) string. Some IE systems require a list of tokens, rather than a string, in which case a tokenizer has to be applied to $text$ to yield a list of strings. However, the tokenizer is extraneous to the definition of a document itself. A corpus is simply a set of documents.

For the purposes of this article, we consider a very simple definition of a schema, namely as a set S of attributes. For the human trafficking domain, the set contains attributes such as phone number, hair color, eye color etc.; in essence, anything that would allow an investigator using these extractions to locate a potentially trafficked victim. More complex schemas and attributes can also be considered (e.g., cluster classes such as *Vendor* explored in Kejriwal and Szekely (2017b)), but will not change the formalism presented herein.

Given an attribute a, we define an information extractor IE_a for that attribute as a system that takes as input the *text* field of a document, and outputs a set of tokens, each of which is denoted as an *instance* of a or equivalently, as having *type a*. The data types of the tokens may be strings, but could also be numbers or dates. Without loss of generality, we assume strings.

Example: Consider the *text* 'Hi, my name is Elsa and I am new in town.' A machine-learning based extractor for the attribute *name* would (ideally) yield the instance *Elsa* when applied to *text* i.e. the extraction *Elsa* would have type *name*.

As the example above indicates, errors in IE can occur for two reasons. First, a correct instance of an attribute may not get extracted by the extractor. Second, an incorrect instance may get extracted. Even here, there are two possibilities. The incorrect instance may be a correct instance of a differently typed extractor e.g., imagine that *Charlotte* got extracted in some sentence as an instance of the city extractor, when in actuality, Charlotte was the name of a person in that sentence. However, it is also quite possible that the wrongly extracted instance is not a correct instance of any type. Some of our research questions in a subsequent section will return to the issue of distinguishing between these different types of 'noise'.

Given an IE IE_a and document D, we can obtain an 'enriched' document (and corpus) by applying IE_a on $D[\,text]$ and obtaining $D' = \left(id, text, \left\{a^1, \ldots, a^m\right\}\right)$, where the third element is the set containing the m instances of attribute a as occurring in D. Similarly, a sequence of IEs $IE_{a_1}, IE_{a_2}, \ldots IE_{a_n}$ can be independently applied for the n attributes $a_1, \ldots, a_n$ in schema S to obtain a fully enriched document D' that records all extracted (from its text) instances of all attributes in its information set.

Given an attribute a, a corpus of enriched documents, and a ground-truth set G of true extractions across all documents in the corpus, we can define accuracy metrics for the extracted instances of a in the corpus. We consider three important metrics in this article, widely used in the IE community, namely precision, recall and F1-measure (the harmonic mean of precision and recall). The precision is the fraction of extractions the IE labeled correctly as positives, while the recall measures how many of the positives in the ground truth the extractor was able to retrieve. Note that each of the metrics can be individually defined for each attribute a, assuming a ground truth G_a of the correct extractions (called the *Positives*) is available. Anything which is not in *Positives* is assumed to be in the set *Negatives*. Clearly, when an IE_a is applied to the corpus, any extraction has to be either in *Positives* or *Negatives*.

In normal situations, these metrics can only be computed and trusted if a good ground truth is available to begin with. Typically this is done by a human annotator who samples some documents and annotates extractions within those documents. This is a laborious process, and much harder to accomplish in the case of sex advertisements, since techniques such as crowdsourcing cannot be effectively leveraged. One of the critical motivations behind this article is to investigate how we can measure (at least in a relative sense) the quality of systems' extractions without access to ground truths. In support of this motivation, we now introduce the simple concept of an attribute extraction network (AEN). The AEN will serve as the central data structure on which empirical studies will be conducted for each attribute.

Definition: An *attribute extraction network* (AEN) N_a is an undirected graph (V, E) where the set of vertices V is defined such that there is one vertex for every *id* in the corpus, and an undirected edge $e_{ij} = \{v_i, v_j\}$ between vertices v_i, v_j $(\in V)$ exists iff a common instance was extracted (for attribute a) for the two documents with IDs corresponding to v_i and v_j.

Figure 2 illustrates an example of how an AEN is defined for five documents (D1-D6) and the *Name* attribute. Each document is a vertex in this representation. The extractions obtained from a given IE system are noted next to the vertex. There is an edge connecting two vertices if their corresponding document representations share an extraction e.g., the extraction 'Mayank' is shared between documents D1 and D2. An important point to note is that a vertex can be a singleton for at least two reasons. First, it may be that no values for attribute a were extracted from the corresponding document. In the figure, document D6 is an example of such a vertex (note that D6 may have extractions for other attributes such as phone number; it just doesn't have an extraction for the attribute *Name* over which this AEN was constructed). Second, it may be that the set of values that were extracted did not get extracted elsewhere (i.e. another document). Thus, by definition that vertex would not be connected to any other vertex in the network. Furthermore, since there is a bijective (1-1) mapping between vertices and documents, we henceforth refer to vertices (also, nodes) as documents for the sake of maintaining a uniform terminology.

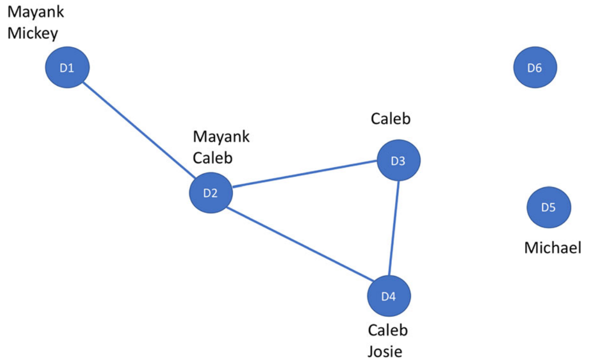


Fig. 2 An example of an Attribute Extraction Network (AEN), assuming the attribute *Name*. Vertices are documents. D6 and D5 are singletons since D5 does not share extractions with any other document, and D6 does not have any extractions

In this article, we refer to a structural network metric as a function that takes an AEN as input and returns either a single point (single-point metric) or a distribution. The only distribution that we will consider in this article is the degree distribution, due to its importance. The structural single-point metrics under consideration are noted in Table 1. Note that the structural network is completely agnostic to what the vertices and edges 'mean' (i.e. their underlying semantics) although, of course, the actual values that a structural network metric would return would depend intimately both on how the network is constructed, and its semantics. One of the goals of this article is to assess the empirical nature and extent of this dependence.

Empirical studies

Earlier in the introduction, we stated two research questions to study the relationship between the network-theoretic metrics presented earlier, and the traditional IE metrics (precision, recall and F-measure). Recall that the first of those questions was based on measurements and comparisons between single-point metrics and the IE metrics, while the second involves similar comparisons but uses an important distributional metric (the degree distribution) rather than the single-point metrics. In this section, we present more details on the data and the empirical methodology for exploring those questions on an online sex trafficking corpus, followed by a report on the results of the analysis.

Table 1 Single point network-theoretic structural metrics considered in this article for some of the empirical studies

Order
Power-law exponent
Number of connected components
Clustering coefficient
Degree correlation
Order (of largest connected component)
Algebraic connectivity
Vertex connectivity
Edge connectivity
Diameter
Average shortest path length

Data

To validate whether network science can be used to assess changes in IE quality without access to a ground truth, we test our hypotheses on IE extractions for which a reference ground truth *is* available. Below, we describe these datasets and the ground-truth in more detail. All datasets were constructed over a large corpus of online sex advertisements that were crawled from (the now shut-down) backpage.com portal during the calendar year of March 2016-2017.

We note that the corpus was collected by an independent contractor funded under the DARPA MEMEX program (mentioned earlier in the *Related Work*), which minimizes chances of dataset bias. The ground-truths were constructed semi-automatically by an academic group of social and political science experts in human trafficking who were not affiliated with the program during ground-truth construction. This ground-truth construction procedure is described in more detail below. The raw HTML pages had to undergo multiple steps of preprocessing and extraction before networks could be constructed. Technical details on webpage preprocessing were provided in our earlier work on information extraction and indicator mining (Kejriwal and Szekely 2017a; Kejriwal et al. 2017). A succinct summary of the datasets is provided in Table 2; further details are provided below.

Ground-truths

In total, the corpus under consideration consists of 11,530 webpages. Multiple domain-specific attributes were extracted from this corpus, including *City, Name, Phone, Address, Service Type*, and even physical attributes such as *Hair Color* and *Eye Color*. In other illicit domains that we have studied (including securities fraud, narcotics, illegal weapons sales online and counterfeit electronics sales), the first three of these were found to be always present in the domain-specific schema that investigators defined. In contrast, *Address* and *Service Type* were more rarely defined, while physical attributes seemed to be exclusive to the online sex trafficking domain. *Name* and *City* are also common in non-illicit domains subject to extraction pipelines e.g., both SpaCy[6] and Stanford NER (two influential open-source IEs tuned for non-illicit domains such as newswire) make available pre-trained modules for *Location* and *Person*, which can be re-normalized to *City* and *Name* as we have considered them in this article (Finkel and Manning 2009).

In keeping with these observations, and to ensure that our findings are relatively generalizable, we consider *Name*, *City* and *Phone* extractions obtained from the corpus. We were provided a ground-truth set of extractions for each document in this corpus, and for each attribute, by an independent group of domain experts and social scientists who had

Table 2 IE datasets (constructed per attribute) used in the empirical studies in this article

Dataset	Explanation
Set1	Ground truth obtained from independent, highly-tuned extractors
Set2	Result of precision-favoring extractions
Set3	Result of recall-favoring extractions
Set4	Set1 ∩ Set2
Set5	Set1 ∩ Set3
Set6	Set1 ∪ Set2
Set7	Set1 ∪ Set3
Set8	Set1 + Random Noise

developed highly tuned rule-based extractors (for all of these three attributes) specifically for sex ads in backpage.com. Typically, such extractors try to encode domain knowledge using unions of regular expressions, followed by post-processing checks. For example, the phone extractor would try to match a sequence of either ten or eleven digits in the ad text, with multiple rules accounting for such obfuscations as word representations of numbers (e.g., one instead of 1), the substitution of o for 0, and so on. Because the focus was on extracting US phones, a post-processing step was to check that the extracted number either had 10 digits, or started with 1 if it had 11 digits. Another step was to remove leading 0s. Checks were also run for numbers that were known to be spammy and occasionally present in ads (e.g., a sequence of nine consecutive 9s or 1s). Similarly, extractors for names and cities also used rules, but additionally, relied on glossaries such as Geonames (Wick and Vatant 2012).

To verify quality, we randomly sampled a set of fifty web documents from the corpus to verify that misclassification rates were low for all attributes. Thus, this dataset can be used in lieu of an exhaustively labeled ground truth, which is not feasible to construct both because of its scale and its real-world qualities. In Table 2, we refer to this dataset as *Set 1*.

Extraction datasets

To test our hypotheses using measures of IE accuracy that are predominantly used in the community (especially precision and recall), we considered two different extraction systems, one of which is precision-favoring and the other of which is recall-favoring. In Table 2, the outputs from these systems are denoted as Sets 2 and 3 respectively.

In illicit domains, structured attributes such as name and phone number are not present in 'infobox' style layouts, but are typically embedded in the text, often in an obfuscated format. This is to avoid direct investigative lookup of an advertiser's street name and contact details using a search engine such as Google. Therefore, in order to extract attributes such as the ones considered in this work, one must first extract the free text from the webpage, following which NLP-centric extraction techniques can be applied on the extracted (and pre-processed) free text. Text scraping from websites is itself a hard problem (due to presence of inserted ad markup, dynamic changes, link structures and variability). We used the Readability Text Extractor (RTE), currently available as the Mercury API[7], to perform the text scraping. We tuned RTE in two different modes. The first mode, which is recall-friendly, is more aggressive and scrapes much of the relevant text, but may also scrape irrelevant text and markup with it. The second mode, which is precision-friendly, tends to be 'cleaner' in that almost all content is relevant, but may miss relevant sentences, especially if there are gaps or links between the relevant portions.

Next, we run identical extraction programs for all three attributes on the precision-friendly and recall-friendly RTE outputs. City and name extractions are obtained using a dictionary based extractor, using existing sets of popular cities and names from a manually curated subset of the GeoNames knowledge base (Wick and Vatant 2012). However, for phones, we used different programs for extracting precision-favoring and recall-favoring phones, since our phone extractor (which has to deal with obfuscation) is based on rules. The precision-favoring phone extractor is applied to the precision-favoring RTE output, and similarly for obtaining recall-favoring phone extractions.

Using the ground-truth dataset (Set 1) and the precision-favoring and recall-favoring datasets (Sets 2 and 3), we can construct other IE datasets expressing varying tradeoffs between noise and quality metrics. We create four new datasets by combining these existing datasets in various ways (e.g., by taking their union). Details on this construction (Sets 4-7) are succinctly formulated in Table 2.

Finally, we also created a synthetic dataset (Set 8) by adding random noise to the ground truth (Set 1) such that quality metrics coincided with those of Set 2. Using the precision and recall values from Set 2 and the number of actual extractions from Set 1, the desired true positive, false positive and false negative values were calculated.

Specifically, for creating more false-positives, a two-step procedure was iteratively employed: (1) a document was chosen at random, and (2) a false extraction, randomly chosen from the dictionary of all extractions observed in the corpus, was added to the extraction set for that document. Similarly, for creating more false-negatives, randomly chosen true extractions are removed from randomly chosen documents in an iterative two-step procedure. Iteration continues till the precision and recall of the constructed dataset equal those of Set 2. Since the number of true positives is fixed, and we are able to precisely control the numbers of false-positives and false-negatives, precision and recall can both be decreased in a controlled and unbiased way. The reason for constructing this dataset is that it proves especially important in assessing some of the results against a reference of random noise, since it allows us to consider whether our real-world IE systems exhibit similar characteristics.

Tables 3, 4 and 5 show some key statistics about all the constructed sets. Since Set 1 is considered as our reference set, we consider it to have perfect quality (1.0 on all quality metrics). In keeping with our intuitions, we find that (relative to Set 3) Set 2 tends to have higher precision (+18-39%), while Set 3 has higher recall (+0.45%-17.8%), though the increase in recall of Set 3 is significantly more diminished by the loss in precision, leading to considerably lower F-scores. Sets 4-8 express a range of tradeoffs; for example, Set 4, which considers the intersection of the ground-truth with an already precision-favoring Set 2, yields perfect precision but at the same level of recall as Set 4. These datasets allow us to counterfactually investigate the different effects of precision and recall on network-theoretic metrics, since they control for one metric.

Finally, as explained earlier, Set 8 was synthetically created by adding random noise to Set 1, such that the quality metrics coincided with those of Set 2; hence, the two sets have expectedly near-identical quality metrics. This also illustrates that, even if a ground-truth were available (such as Set 1) to a practitioner, she would not be able to distinguish

Table 3 Dataset characteristics for city extractions

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
No. of unique extractions	328	1,685	1,489	253	253	1,760	1,563	1,865
No. of extractions per advertisement	1.0	3.4720	17.1507	0.7862	0.7818	3.6858	17.3689	3.4722
No. of ads with no extractions	0	64	65	2,465	2,516	0	0	178
No. of ads with at least 1 extraction	11,530	11,466	11,465	9,065	9,014	11,530	11,530	11,352
Precision	1.0	0.2264	0.0456	1.0	1.0	0.2713	0.0576	0.2264
Recall	1.0	0.7862	0.7818	0.7862	0.7818	1.0	1.0	0.7862
F-score	1.0	0.3516	0.0861	0.8803	0.8775	0.4268	0.1089	0.3516

Table 4 Dataset characteristics for name extractions

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
No. of unique extractions	925	631	900	468	594	1,076	1,218	1,169
No. of extractions per advertisement	0.7563	0.6271	2.7060	0.3794	0.5140	1.0039	2.9484	0.6269
No. of ads with no extractions	4,708	5,937	532	7,607	6,536	3,613	328	6,362
No. of ads with at least 1 extraction	6,822	5,593	10,998	3,923	4,994	7,917	11,202	5,168
Precision	1.0	0.6050	0.1899	1.0	1.0	0.7533	0.2565	0.6050
Recall	1.0	0.5016	0.6796	0.5016	0.6796	1.0	1.0	0.5015
F-score	1.0	0.5485	0.2969	0.6681	0.8092	0.8593	0.4083	0.5484

between Sets 2 and 8 based only on IE metrics. We show subsequently, however, that the structural properties of the extraction graphs of Sets 2 and 8 markedly differ.

Experiments and methods

To answer the first research question (RQ 1), we devised a set of quantitative experimental methods to record the variance in structural metrics for each of the eight datasets listed in Table 2. Note that structural metrics are unsupervised, requiring mechanical computations that depend only on the structure of the network. For each of the three attributes under consideration, we compute the individual Pearson correlation between the precision, recall and F-score, and several well-known network-theoretic structural metrics such as described in *Formalism* using eight single-point measurements for the correlations (one data point per metric per dataset in Table 2). We do not consider non-single-point metrics such as the degree distribution, since its investigation falls specifically within the purview of RQ 2. Because the eight datasets in Table 2 have varying qualities on the different IE metrics of interest to us (precision, recall and F-measure), consistent changes in structure across all three attributes enable us to take a principled approach to answering RQ 1.

Furthermore, with a view to assessing if noise in real-world IE systems exhibits significantly non-random tendencies, we report the same structural metrics that we considered in the methodology above for each set and attribute in Table 2, and study the specific differences between Set 8 and Set 2, since Set 8 has the same accuracy as Set 8, but with noise inserted randomly.

Table 5 Dataset characteristics for phone extractions

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
No. of unique extractions	2,298	1,756	3,668	1,697	1,716	2,355	4,248	2,206
No. of extractions per advertisement	1.0850	0.9639	1.3004	0.9490	0.9649	1.0999	1.4206	0.9638
No. of ads with no extractions	194	651	367	708	531	181	132	1,623
No. of ads with at least 1 extraction	11,336	10,879	11,163	10,822	10,999	11,349	11,398	9,907
Precision	1.0	0.9845	0.7419	1.0	1.0	0.9864	0.7638	0.9845
Recall	1.0	0.8746	0.8892	0.8746	0.6796	1.0	1.0	0.8745
F-score	1.0	0.9263	0.8089	0.9414	0.8092	0.9931	0.8661	0.9263

The methodology for exploring RQ 2 is fairly straightforward; we compute and plot the degree distribution of the extraction networks underlying the sets in Table 2. We also refer to the power-law coefficient of each network, computed earlier in the data collected for answering RQ 2, to assess to what extent each distribution obeys the power law. We also study how the power law distribution for each network evolves for a given attribute as performance changes gradually (across the spectrum from Set 1 to Set 8).

Results

We report results for both research questions enumerated earlier. Each research question (RQ) is considered individually below.

Research question 1

Recall that the first research question involved detecting patterns in structural metrics' changes with changes in IE quality. We note the primary observations that emerged from conducting the RQ 1 experiments, using the methodology described earlier, below:

- First, precision was found to be strongly correlated with several structural metrics, as quantified in Table 6, which records the Pearson correlation coefficient using the 8-point precision vector of the eight datasets in Table 2, and the corresponding values of the single-point structural metrics computed over their respective extraction networks. In some cases, the correlations seem intuitive and even obvious. For example, the relatively strong negative correlation between *Order* and precision can be explained as follows. Since the order corresponds to the set of all entities (for an attribute) extracted over the entire set of documents, and since every unique entity has, in practice, some non-zero probability of being noise, networks with a higher order tend to have lower precision. This is especially true when an attribute in question contains entities from some pre-specified 'universal' set, which is true for names[8] and cities. In contrast, phones, which are syntactically constrained, but tend to accommodate many more possible unique values, show a weaker (but still quite strong) negative correlation.
- Second, more interestingly, the 'erroneous' edges in less precise extraction networks tend to serve as 'weak ties' that end up collapsing two or more connected components into a single connected component, reducing the number of connected components. In other words, less precise edges tend to straddle components (a rough definition of what would constitute a weak tie in network science). This suggests a potential line of attack in cleaning up noisy extractions, by exploiting hierarchical or agglomerative clustering algorithms (Murtagh and Legendre 2014) that may be able to detect such weak ties (e.g., by iteratively breaking up connected components into clusters using mechanisms such as betweenness centrality for assigning weights to edges). The empirical utility of such methods is an important agenda that we will pursue in future work.
- Third, precision is positively correlated with the *Clustering Coefficient* of the extraction network, but the correlation is not as strong as between precision and the number of connected components. This implies that cleaner extraction sets yield a smaller number of, but more tightly knit, groups (in the underlying extraction network) as compared with noisier extraction sets. In other words, in aggregate, the incorrect extraction edges tend to contribute to non-transitivity, since clustering

Table 6 Pearson correlation coefficients between precision and network metrics

Network metric	City	Name	Phone
Order	-0.7292*	-0.8219*	-0.4537
Power-law exponent	-0.6995	0.8628**	-0.6169
Num. connected components	0.9741**	0.9624**	0.9886**
Clustering coefficient	0.7010	0.9017**	0.8362**
Degree correlation	0.3245	0.2297	0.157
Order (of largest connected component)	-0.9839**	-0.7776*	-0.9958**
Algebraic connectivity	0.9837**	-0.5925	0.7055
Vertex connectivity	0.9838**	0.5204	0.6025
Edge connectivity	0.9834**	0.5204	0.7229*
Diameter	-0.5471	0.8368**	-0.9950**
Avg. shortest path length	-0.1373	0.4637	-0.9964**

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component. * and ** respectively indicate statistical significance at the 95% and 99% level for the two-tailed test for the correlation coefficient with null hypothesis $\rho = 0$

coefficient is related to the number of triadic closures (indicating high transitivity) in the network. On average, therefore, given two links $n_1 - n_2$ and $n_1 - n_3$, all else being equal, a third link $n_2 - n_3$ introduced by a real-world extraction system is more likely to be correct than in the absence of either $n_1 - n_2$ or $n_1 - n_3$ (or both). This suggests another potential line of attack in trying to clean up noisy extractions (or selecting a system under the expectation of high precision without access to ground-truth) by systematically making use of global information.

- Fourth, precision is also positively correlated with the Vertex and Edge Connectivity of the largest connected component. Adding more incorrect extraction edges leads to lower connectivity in the larger connected component compared to the original closely connected component. This offers a finer-grained 'check' on systems' precision, as opposed to a coarse-grained classification (of whether a given extraction set is more precise than another extraction set) compared to the previous observation, which would only check the size of the largest component.
- Interestingly, in contrast with precision, recall is not heavily correlated with any of the metrics described above, whether positively or negatively. Table 7 shows the correlation between recall and the various metrics discussed above in the context of precision.
- Finally, because F-score is the harmonic mean of both precision and recall, it was unsurprisingly found to be correlated positively with the number of connected components, and also the clustering coefficient, of the network. The correlations were not as strong as those of precision (Table 8).

Brief Summary. The results show that certain structural metrics are excellent predictors of the overall performance of an extraction system, especially if precision is of interest. In contrast, recall cannot be predicted very accurately. We note that there is also variance between the attributes, though not as strong as one might expect, given that they are very different from one another. In general, we found that, in terms of evaluating RQ1, the *Name* attribute tended to be more heterogeneous and less predictive compared to *City* and *Phone* attributes. Other possible limitations of the study are described in the "Discussion" section.

Table 7 Pearson correlation coefficients between recall and network metrics

Network metric	City	Name	Phone
Order	0.4816	0.5794	0.3993
Power-law exponent	-0.06280	-0.1790	0.3877
Num. connected components	0.0228	-0.2304	-0.1901
Clustering coefficient	0.1563	-0.1438	-0.4472
Degree correlation	-0.0443	-0.4175	-0.6218
Order (of largest connected component)	0.0752	0.4372	0.2662
Algebraic connectivity	-0.0650	0.0182	0.1217
Vertex connectivity	-0.0648	-0.4629	0.3248
Edge connectivity	-0.0660	-0.4629	-0.0360
Diameter	-0.1662	-0.5824	0.2573
Avg. shortest path length	-0.0277	-0.7176*	0.2519

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component. * indicates statistical significance at the 95% level for the two-tailed test for the correlation coefficient with null hypothesis $\rho = 0$

Random vs. non-random noise

Recall that one of our goals had been to study the properties of real-world extraction noise; namely, is the noise random? We proposed studying these properties by first introducing attribute-specific random noise in the ground-truth network (Set 8 in Table 2) till it had the same precision, recall and F-scores as Set 2 (Tables 3, 4 and 5) for that attribute. Using *only* the accuracy metrics, there is no difference, in aggregate, between Sets 2 and 8. However, our (subsequently described) observations show that there are considerable structural differences between the two networks, providing evidence that the noise incorporated in real-world extraction settings is indeed significantly non-random. Furthermore, in deviating from randomness, the noise exhibits some clear patterns, lending credence to the observations and summary in the previous section as well.

First, we illustrate (in Fig. 3) the degree distribution of the Set 8 (i.e. random noise) network and compare it to the Set 2 network (obtained from a real-world 'high precision' extraction system). The figure reveals that the number of lower-degree nodes tend to be higher in the Set 8 network, as would be expected with random noise, while the number of higher-degree nodes tend to be higher in the Set 2 network. In investigating RQ 1, we

Table 8 Pearson correlation coefficients between F-Score and network metrics

Network metric	City	Name	Phone
Order	-0.6284	-0.5268	-0.0769
Power-law Exponent	-0.5918	0.7635*	-0.1704
Num. Connected Components	0.9484**	0.8211*	0.6347
Clustering Coefficient	0.6395	0.7352*	0.3150
Degree Correlation	0.2693	0.0863	-0.3799
Order (of Largest Connected Component)	-0.9401**	-0.3578	-0.5885
Algebraic Connectivity	0.9413**	-0.7014	0.6583
Vertex Connectivity	0.9415**	0.3980	0.7384*
Edge Connectivity	0.9407**	0.3980	0.5411
Diameter	-0.4459	0.3831	-0.5926
Avg. Shortest Path Length	-0.3948	-0.1305	-0.5984

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component. * and ** respectively indicate statistical significance at the 95% and 99% level for the two-tailed test for the correlation coefficient with null hypothesis $\rho = 0$

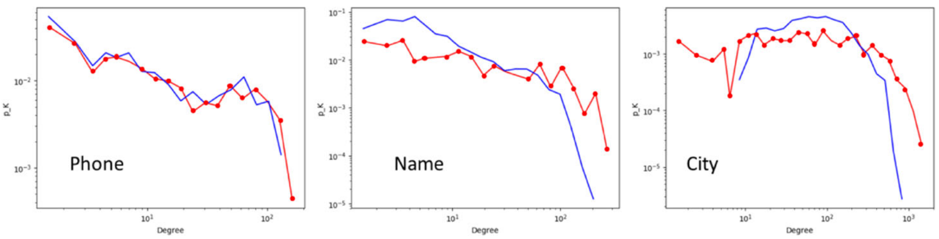


Fig. 3 (Log-log) Degree distributions of Set 2 (Dotted Red) and Set 8 (Solid Blue) extraction networks for all three attributes. The Y-axis records the empirical probability of each degree value

saw earlier that noise in real-world networks can have a 'weak link' effect in that noisy links end up connecting otherwise disconnected components than might be expected by chance. The figure agrees with this intuition, in that higher-degree nodes continue to increase in degree (thereby seeming to obey scale-free assumptions) with addition in real-world noise, in contrast with random noise that skews the degree distribution in the reverse direction.

Second, the clustering coefficient of the Set 8 network is consistently lower than that of the Set 2 network (see Tables 9, 10, and 11), providing more evidence that real-world erroneous extractions are more localized than would be true for random errors. Not only that, but the same errors seem to recur consistently across documents, which leads to their clustering by means of common (error-prone) extractions. We also note that the largest connected component of the Set 2 network has a smaller diameter for two out of the three attributes (Phone and Name) compared to the random network. As is true for other real-world networks exhibiting (approximately) power-law degree distributions (such as social networks), a real-world noise network also tends to exhibit 'small world' properties (in comparison with random networks).

Brief Summary. Real-world extraction systems are not noisy in random ways, which (arguably) provides a compelling reason for using network science in the first place for studying their noise. More practically, it explains why active learning approaches lead to super-linear (with respect to labeling effort) gains when properly used, since the same

Table 9 Single-point structural network metrics for *City* extraction datasets

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
Order	11492	11460	11465	9039	8988	11525	11530	11352
Power-law exponent	0.6431	2.2155	1.3639	0.5924	0.5889	2.3504	1.3709	3.1598
Num. connected components	290	11	2	227	227	7	2	1
Clustering coefficient	0.9944	0.8387	0.8377	0.9947	0.9947	0.8367	0.8367	0.5493
Degree correlation	1.0	0.5641	1.0	1.0	0.5542	0.5542	0.6397	0.6275
Order (of largest connected component)	533	11430	11463	533	513	11510	11528	11352
Algebraic connectivity	532.9999	0.8757	0.4989	533.0	513.0	0.7767	FAIL	7.8806
Vertex connectivity	532	1	3	532	512	1	FAIL	8
Edge connectivity	532	1	3	532	512	1	FAIL	8
Diameter	1	7	5	1	1	7	FAIL	4
Avg. shortest path length	1	3.4945	1.9510	1	1	2.6403	FAIL	2.3603

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component. FAIL indicates that the metric could not be computed for that set, usually due to a program timeout or memory issue

Table 10 Single-point structural network metrics for *Name* extraction datasets

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
Order	6603	5429	10987	3776	4833	7716	11185	5095
Power-law exponent	2.7054	1.5843	0.8963	2.4881	2.5301	2.1872	0.8970	2.8074
Num. connected components	195	153	3	193	201	147	4	138
Clustering coefficient	0.9273	0.9162	0.8439	0.9440	0.9355	0.8986	0.8410	0.8502
Degree correlation	0.7681	0.7612	0.9803	0.7651	0.8203	0.6552	0.3280	0.6340
Order (of largest connected component)	4859	4376	8406	1829	2892	6842	11005	4622
Algebraic connectivity	0.0595	0.0842	0.9967	0.0123	0.0294	0.1280	0.9989	0.0738
Vertex connectivity	1	1	1	1	1	1	1	1
Edge connectivity	1	1	1	1	1	1	1	1
Diameter	11	11	5	17	15	9	4	13
Avg. shortest path length	0	4.0801	1.7173	5.7881	5.3785	3.8888	5.1463	4.7180

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component

errors seem to be occurring 'independently' in multiple documents (Thompson et al. 1999). If the noise had been truly random, active learning would be much less effective, since there would be a higher probability of sampling lower-degree nodes in Fig. 3.

Research question 2

Finally, in investigating the dependency of the degree distribution on the varying levels of noise and precision-recall tradeoff, we present degree distribution (log-log) plots in Figs. 4, 5 and 6. Contrary to social networks and other 'natural' networks, noise in the extraction networks does not follow a clear power law. However, with the exception of the phone extraction network's degree distributions, which are fairly uniform across all sets (an artifact that may be the result of phone extractions being of generally higher quality than other extractions[9]), we note that the networks for Sets 3 and 7 are most erratic (compared to the Set 1 ground truth network). Considering the data in Tables 3 and 4, we find that these are the two sets with the lowest F-scores. In contrast, Sets 5 and 6 for

Table 11 Single-point structural network metrics for *Phone* extraction datasets

Metric	Set1	Set2	Set3	Set4	Set5	Set6	Set7	Set8
Order	10286	9863	10252	9817	9993	10313	10488	8989
Power-law exponent	0.9887	1.5696	1.5910	0.9975	1.0085	1.5389	1.5800	1.3138
Num. connected components	656	610	350	619	633	650	368	589
Clustering coefficient	0.9476	0.9547	0.9314	0.9563	0.9551	0.9463	0.9273	0.9370
Degree correlation	0.9816	0.9779	0.9803	0.9869	0.9865	0.9730	0.9788	0.9780
Order (of largest connected component)	189	189	8406	189	189	189	8636	199
Algebraic connectivity	2.9177	2.9730	0.0046	1.9555	1.9570	3.9630	0.0047	0.2295
Vertex connectivity	4	3	1	2	2	4	1	1
Edge connectivity	3	3	1	3	3	4	1	1
Diameter	3	3	19	3	3	3	19	5
Avg. shortest path length	1.6318	1.5709	6.9405	1.6518	1.6417	1.5537	6.8760	1.9739

Since the network is disconnected, all metrics listed below *Order (of Largest Connected Component)* are computed over the largest connected component

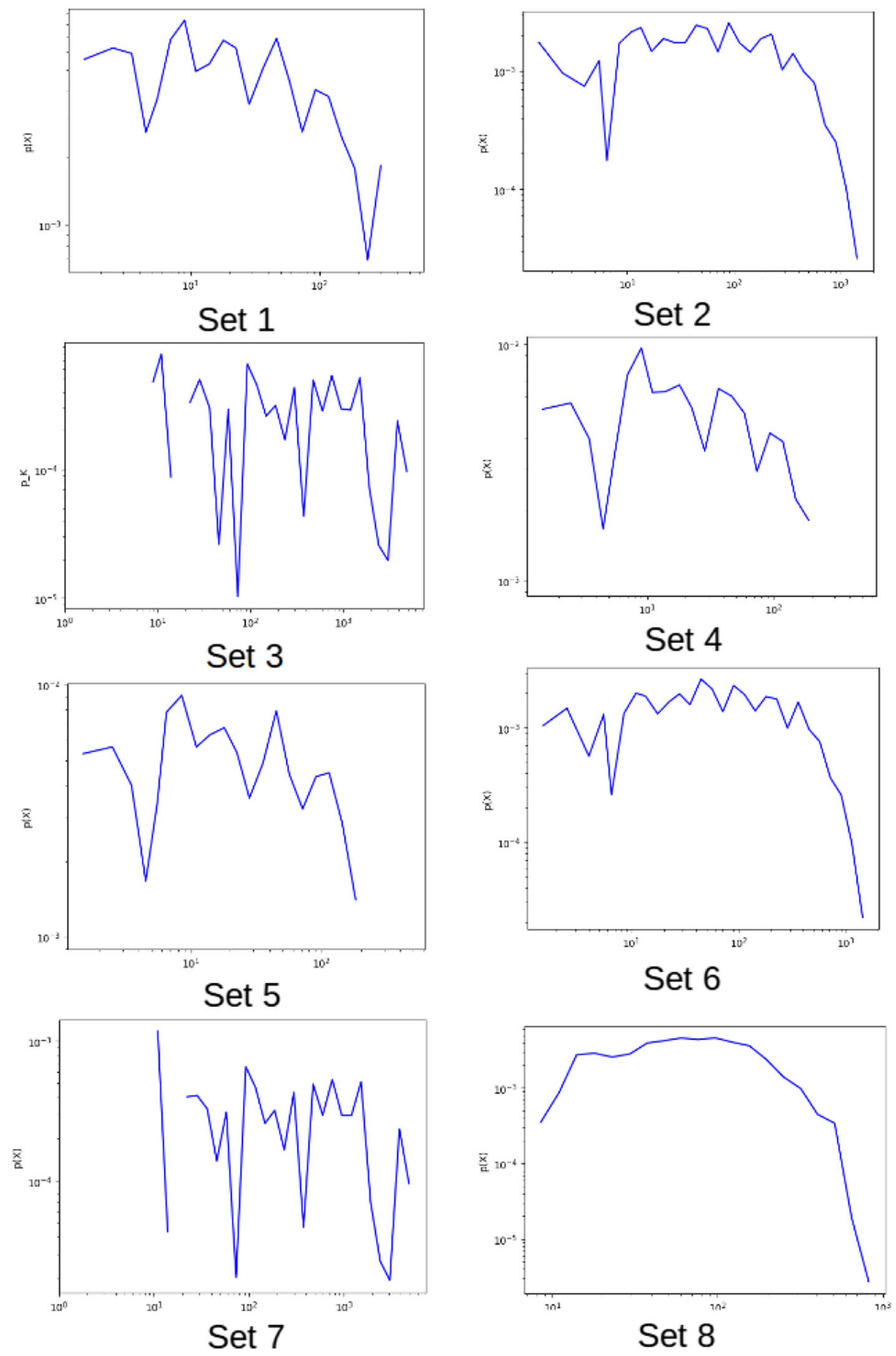


Fig. 4 Degree distributions of the *City* extraction network for the eight datasets described in Table 3, in support of the findings in *Research Question 2*. The X-axis and Y-axis respectively plot the degree and its empirical frequency in the network. Scales are log-log

both attributes have F-scores greater than 80%. Correspondingly, the degree distributions for these sets are much smoother. Although difficult to quantify, the results show that the degree distribution could potentially be used to diagnose whether the F-score of an extraction set is abnormally low.

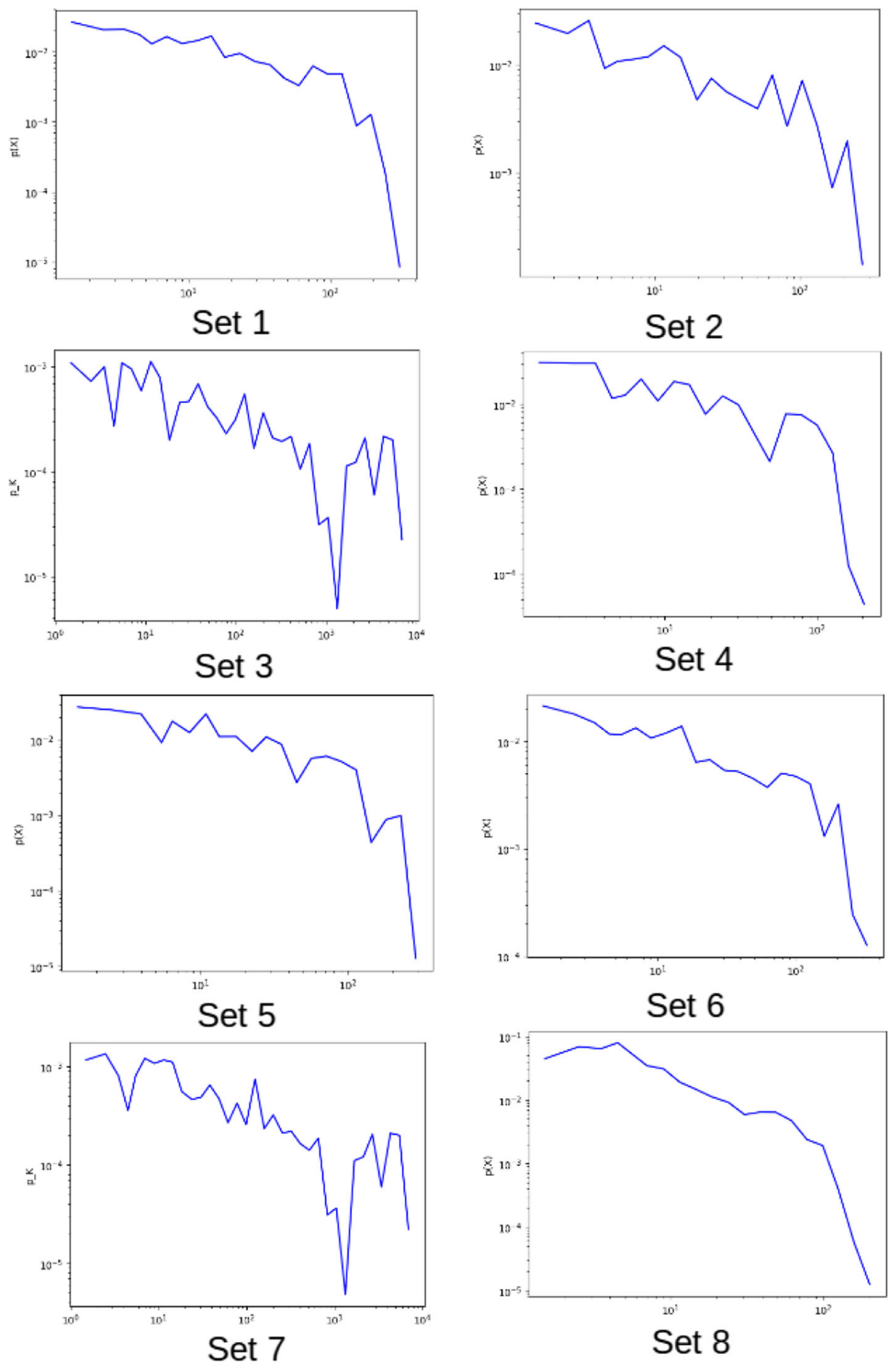


Fig. 5 Degree distributions of the *Phone* extraction network for the eight datasets described in Table 5, in support of the findings in *Research Question 2*. The X-axis and Y-axis respectively plot the degree and its empirical frequency in the network. Scales are log-log

Discussion

In exploring the research questions, we utilized a select set of network metrics on the AEN as an approach for evaluating the quality of rival IE systems without access to ground-truth, in an unusual domain such as human trafficking where ground-truths are difficult

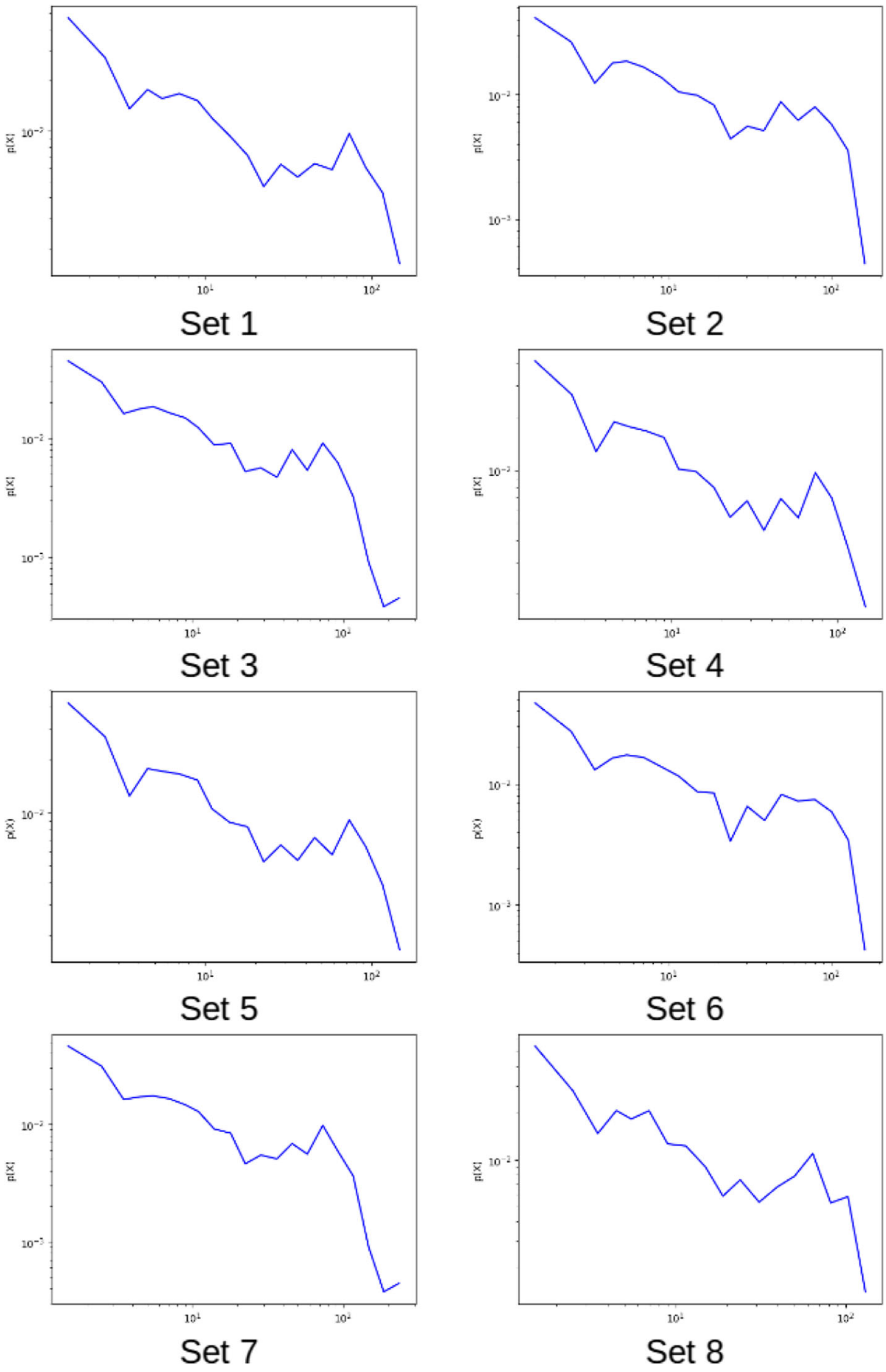


Fig. 6 Degree distributions of the *Name* extraction network for the eight datasets described in Table 4, in support of the findings in *Research Question 2*. The X-axis and Y-axis respectively plot the degree and its empirical frequency in the network. Scales are log-log

to acquire across the Web. These network metrics were found to be correlated with some IE accuracy metrics, particularly precision. Although it is not currently feasible to provide a mathematical justification for why this turned out to be the case (although in future work, we are looking to develop a theoretical model explaining this finding), we posit two intuitive reasons below:

1 Noise in IE is non-random i.e. if a word or phrase got mis-extracted in one document, there is a higher-than-normal probability that it will get mis-extracted in another document. This occurs despite the observations that both documents were generated independently, the contexts surrounding the word are distinct in both documents, and the training data was sufficiently representative. Intuitively, this occurs because there is an extraneous property that leads to the noise. For example 'Charlotte' may be getting mis-extracted more often than 'Los Angeles' by a Named Entity Recognition system (Nadeau and Sekine 2007), despite representative training data, because Charlotte is also a common name. Charlotte is what a practitioner would define as a 'difficult' example, even though it is hard to formalize what makes one example more difficult than another.

2 In aggregate, noise seems to have macroscopic structure that can be formally quantified using concepts from network science. One reason for this is that the universe of extractions (i.e. all possible extractions) tends to be bounded in practice. For example, there is a finite number of locations in the world, and although any string could potentially be a name, the number of names in a corpus tends to be bounded. Because extractions (and also mis-extractions) are repeated across documents, certain regular structures and patterns may emerge. Consider again the case of 'Charlotte' statistically: assuming it gets incorrectly extracted by a recall-friendly system, along with the true city extraction, it is statistically unlikely that the true extraction will also be a city in North Carolina. In the broader AI community, probabilistic techniques (such as Probabilistic Soft Logic or PSL) have exploited this observation to ingest extractions from multiple independent IEs, and identify true extractions by probabilistically reasoning about such patterns (Kimmig et al. 2012). However, PSL needs clear domain rules (which is beyond reach of non-technical investigative experts), and knowledge graph identification systems that rely on PSL try to combine multiple IE and Entity Resolution systems to leverage such statistical knowledge.

We also note that the study is not without its limitations, which must be borne in mind before applying the findings to other HT datasets, or to datasets from similar illicit (or even non-illicit domains). Importantly, while the systems and datasets considered in this article are real-world, the structural metrics are 'global', meaning that, in general, it is considerably more difficult to predict precisely when a given system is wrong (i.e. pinpoint individual wrong extractions or links). In the future, it may be possible to use the concepts developed in this work in a machine learning setting to make such microscopic predictions; however, at present, the network metrics are computed over the full network rather than on a per-node or per-edge basis. However, because the network metrics capture an aggregate property of the performance of the underlying IE system (e.g., whether it is precision-favoring or not), they could be used to configure the IE system through hyperparameter optimization (Bergstra et al. 2011; Eggensperger et al. 2013). Intuitively, each set of hyperparameters yields a 'different' IE system, expressing a performance tradeoff typically captured through ROC curves (plotted using a validation set) in the machine learning literature. However, the network metrics are computed in an unsupervised fashion, and do not need labeled data.

Conclusion

In this article, we addressed the problem of assessing and profiling data quality in competing information extraction systems over domains that are unusual, have no ground truth annotations, but are consequential in the real world. We conducted a detailed empirical study using extractions covering three attributes, and different IE precision-recall tradeoffs, over a large corpus of webpages in the sex advertisement domain. The empirical studies illustrate some interesting aspects of noise in IE systems. For example, we found that, in real-world extraction systems, edges introduced in the attribute extraction network due to erroneous extractions tend to be of the 'weak tie' variety and lead to larger connected components. Recall was not found to exhibit strong dependencies on any structural metrics. Finally, noise distributions were found to exhibit non-random tendencies, with more predictable patterns emerging for lower levels of noise.

In current ongoing research, we are looking to release a software package that is able to use regression analysis to predict precision, recall and F-Measure scores for different configurations of an IE system, given baseline scores with respect to a default configuration. This package is expected to serve a useful purpose both in active learning and for determining system improvement with small or no ground truths.

Endnotes

[1] Although state-of-the-art IE is still supervised, unsupervised approaches have come a long way (Nadeau and Sekine 2007).

[2] Now shut down and under investigation by federal authorities.

[3] This term is widely regarded as being a misnomer by many practitioners in NLP, since the data does have structure, although it is not parseable (with guaranteed high quality) by machines.

[4] https://www.darpa.mil/program/memex

[5] https://www.scientificamerican.com/article/human-traffickers-caught-on-hidden-internet/

[6] https://spacy.io/

[7] https://mercury.postlight.com/web-parser/

[8] Most name extractions in our corpus could be derived from a broad lexicon of English names.

[9] No phone extraction set has F-score below 80%, while name and city extraction sets exhibit considerably more variety.

Acknowledgements
Not Applicable.

Authors' contributions
Both authors have contributed equally to the ideas in the work presented herein. All experimental work was conducted by Kapoor, while the writing was under Kejriwal's purview. Both authors read and approved the final manuscript.

Authors' information
Not Applicable.

Funding
This research was supported by the Defense Advanced Research Projects Agency (DARPA) and the Air Force Research Laboratory (AFRL) under contract number FA8750- 14-C-0240. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of DARPA, AFRL, or the U.S. Government.

Availability of data and materials
The authors are willing to release the logs of all experimental runs, and also name and city extractions with anonymized advertisement IDs. Phone extractions cannot be released due to privacy constraints, since the phones could still be active. We also cannot release the raw data itself due to the sensitivity of the material.

Competing interests

The authors declare that they have no competing interests.

Received: 16 November 2018 Accepted: 6 June 2019
Published online: 27 June 2019

References

Ahn D (2006) The stages of event extraction. In: Proceedings of the Workshop on Annotating and Reasoning About Time and Events. Association for Computational Linguistics. pp 1–8. https://aclweb.org/anthology/papers/W/W06/W06-0901/

Alvari H, Shakarian P, Snyder JK (2016) A non-parametric learning approach to identify online human trafficking. In: Intelligence and Security Informatics (ISI), 2016 IEEE Conference On. IEEE. pp 133–138. https://ieeexplore.ieee.org/document/7745456

Alvari H, Shakarian P, Snyder JK (2017) Semi-supervised learning for detecting human trafficking. Secur Inform 6(1):1

Banko M, Cafarella MJ, Soderland S, Broadhead M, Etzioni O (2007) Open information extraction from the web. In: IJCAI, vol. 7. IJCAI, hyderabad. pp 2670–2676. https://www.ijcai.org/proceedings/2007

Barabási A-L, et al (2016) Network science. https://www.cambridge.org/us/academic/subjects/physics/statistical-physics/network-science?format=HB

Berger SI, Iyengar R (2009) Network analyses in systems pharmacology. Bioinformatics 25(19):2466–2472

Bergstra JS, Bardenet R, Bengio Y, Kégl B (2011) Algorithms for hyper-parameter optimization. In: Advances in Neural Information Processing Systems. pp 2546–2554. https://papers.nips.cc/paper/4443-algorithms-for-hyper-parameter-optimization

Borgatti SP, Mehra A, Brass DJ, Labianca G (2009) Network analysis in the social sciences. Science 323(5916):892–895

Burbano D, Hernandez-Alvarez M (2017) Identifying human trafficking patterns online. IEEE. https://ieeexplore.ieee.org/document/8247461

Chang C-H, Kayed M, Girgis MR, Shaalan KF (2006a) A survey of web information extraction systems. IEEE Trans Knowl Data Eng 18(10):1411–1428

Chang C-H, Kayed M, Girgis MR, Shaalan KF (2006b) A survey of web information extraction systems. IEEE Trans Knowl Data Eng 18(10):1411–1428

Chen H (2011) Dark web: Exploring and data mining the dark side of the web. https://link.springer.com/chapter/10.1007/978-3-642-29892-9_1

Chen P, Redner S (2010) Community structure of the physical review citation network. J Informetrics 4(3):278–290

Chinchor N (1998) Overview of muc-7. In: Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998. https://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_proceedings/overview.html

Doan A, Halevy A, Ives Z (2012) Principles of data integration. https://www.sciencedirect.com/book/9780124160446/principles-of-data-integration

Easley D, Kleinberg J, et al (2010) Networks, Crowds, and Markets vol. 8. Cambridge University Press Cambridge. https://www.cambridge.org/us/academic/subjects/computer-science/algorithmics-complexity-computer-algebra-and-computational-g/networks-crowds-and-markets-reasoning-about-highly-connected-world?format=HB

Eggensperger K, Feurer M, Hutter F, Bergstra J, Snoek J, Hoos H, Leyton-Brown K (2013) Towards an empirical foundation for assessing bayesian optimization of hyperparameters. In: NIPS Workshop on Bayesian Optimization in Theory and Practice, vol. 10. NIPS, Lake Tahoe. p 3. https://nips.cc/Conferences/2013

Elmagarmid AK, Ipeirotis PG, Verykios VS (2007) Duplicate record detection: A survey. Knowl Data Eng IEEE Trans 19(1):1–16

Etzioni O, Banko M, Soderland S, Weld DS (2008) Open information extraction from the web. Commun ACM 51(12):68–74

Finkel JR, Manning CD (2009) Joint parsing and named entity recognition. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics. pp 326–334. https://aclweb.org/anthology/papers/N/N09/N09-1037/

Freitag D (2000) Machine learning for information extraction in informal domains. Mach Learn 39(2-3):169–202

Gavin A-C, Bösche M, Krause R, Grandi P, Marzioch M, Bauer A, Schultz J, Rick JM, Michon A-M, Cruciat C-M, et al (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415(6868):141

Greenberg SA (2009) How citation distortions create unfounded authority: analysis of a citation network. Bmj 339:2680

Hultgren M, Jennex ME, Persano J, Ornatowski C (2016) Using knowledge management to assist in identifying human sex trafficking. In: System Sciences (HICSS), 2016 49th Hawaii International Conference On. IEEE. pp 4344–4353. https://ieeexplore.ieee.org/document/7427725

Hultgren M, Whitney J, Jennex ME, Elkins A (2018) A knowledge management approach to identify victims of human sex trafficking. CAIS 42:23

Hummon NP, Dereian P (1989) Connectivity in a citation network: The development of dna theory. Soc Netw 11(1):39–63

Kapoor R, Kejriwal M, Szekely P (2017) Using contexts and constraints for improved geotagging of human trafficking webpages. arXiv preprint arXiv:1704.05569

Kejriwal M, Szekely P (2017a) Information extraction in illicit web domains. In: Proceedings of the 26th International Conference on World Wide Web. International World Wide Web Conferences Steering Committee. pp 997–1006. https://dl.acm.org/citation.cfm?id=3052642

Kejriwal M, Szekely P (2017b) Knowledge graphs for social good: an entity-centric search engine for the human trafficking domain. IEEE Trans Big Data 1:1–1

Kejriwal M, Szekely P, Knoblock C (2018) Investigative knowledge discovery for combating illicit activities. IEEE Intell Syst 1:53–63

Kejriwal M, Ding J, Shao R, Kumar A, Szekely P (2017) Flagit: A system for minimally supervised human trafficking indicator mining. arXiv preprint arXiv:1712.03086

Kimmig A, Bach S, Broecheler M, Huang B, Getoor L (2012) A short introduction to probabilistic soft logic. In: Proceedings of the NIPS Workshop on Probabilistic Programming: Foundations and Applications. NIPS, Lake tahoe. pp 1–4. https://nips.cc/Conferences/2012

Kleinberg JM (2007) Challenges in mining social network data: processes, privacy, and paradoxes. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp 4–5. https://dl.acm.org/citation.cfm?id=1281195

Knoke D, Yang S (2008) Social Network Analysis vol. 154. Sage. https://us.sagepub.com/en-us/nam/social-network-analysis/book228826

Kushmerick N, Weld DS, Doorenbos R (1997) Wrapper induction for information extraction:729–37. Washington: University of Washington

Lerman K, Minton S, Knoblock CA (2003) Wrapper maintenance: A machine learning approach. J Artif Intell Res(JAIR) 18:149–181

Leskovec J, Huttenlocher D, Kleinberg J (2010) Predicting positive and negative links in online social networks. In: Proceedings of the 19th International Conference on World Wide Web. ACM. pp 641–650. https://dl.acm.org/citation.cfm?id=1772756

Li X, Chen H, Huang Z, Roco MC (2007) Patent citation network in nanotechnology (1976–2004). J Nanoparticle Res 9(3):337–352

Moreno JL (1946) Sociogram and sociomatrix. Sociometry 9:348–349

Murtagh F, Legendre P (2014) Ward's hierarchical agglomerative clustering method: which algorithms implement ward's criterion?. J Classif 31(3):274–295

Muslea I, Minton S, Knoblock C (1998) Stalker: Learning extraction rules for semistructured, web-based information sources. In: Proceedings of AAAI-98 Workshop on AI and Information Integration. AAAI Press Menlo Park, CA. pp 74–81. https://www.aaai.org/Library/Workshops/ws98-14.php

Nadeau D, Sekine S (2007) A survey of named entity recognition and classification. Lingvisticae Investigationes 30(1):3–26

Niu F, Zhang C, Ré C, Shavlik JW (2012) Deepdive: Web-scale knowledge-base construction using statistical learning and inference. VLDS 12:25–28

Rabbany R, Bayani D, Dubrawski A (2018) Active search of connections for case building and combating human trafficking. In: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM. pp 2120–2129. https://dl.acm.org/citation.cfm?id=3220103

Ritter A, Etzioni O, Clark S, et al (2012) Open domain event extraction from twitter. In: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM. pp 1104–1112. https://dl.acm.org/citation.cfm?id=2339704

Schreiber F, Schwöbbermeyer H (2005) Mavisto: a tool for the exploration of network motifs. Bioinformatics 21(17):3572–3574

Szekely P, Knoblock CA, Slepicka J, Philpot A, Singh A, Yin C, Kapoor D, Natarajan P, Marcu D, Knight K, et al (2015) Building and using a knowledge graph to combat human trafficking. In: International Semantic Web Conference. Springer. pp 205–221. https://link.springer.com/chapter/10.1007/978-3-319-25010-6_12

Tang J, Lou T, Kleinberg J (2012) Inferring social ties across heterogenous networks. In: Proceedings of the Fifth ACM International Conference on Web Search and Data Mining. ACM. pp 743–752. https://dl.acm.org/citation.cfm?id=2124382

Thompson CA, Califf ME, Mooney RJ (1999) Active learning for natural language parsing and information extraction. In: ICML. Citeseer. pp 406–414. https://dl.acm.org/citation.cfm?id=657614

Voorhees EM, et al (1999) The trec-8 question answering track report. In: Trec Vol. 99. pp 77–82

Wasserman S, Faust K (1994) Social Network Analysis: Methods and Applications vol. 8. Cambridge university press. https://www.cambridge.org/us/academic/subjects/sociology/sociology-general-interest/social-network-analysis-methods-and-applications?format=PB

Wernicke S, Rasche F (2006) Fanmod: a tool for fast network motif detection. Bioinformatics 22(9):1152–1153

Whitney J, Jennex M, Elkins A, Frost E (2018) Don't want to get caught? don't say it: The use of emojis in online human sex trafficking ads. https://scholarspace.manoa.hawaii.edu/handle/10125/50426

Wick M, Vatant B (2012) The geonames geographical database. Available from World Wide Web: http://geonames.org Accessed 20 June 2019

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.