Scalable Computing: Practice and Experience
Volume 9, Number 4, pp. 271–279. http://www.scpe.org

ISSN 1895-1767

UNEVENNESS IN NETWORK PROPERTIES ON THE SOCIAL SEMANTIC WEB

RAF GUNS*

Abstract. This paper studies unevenness in network properties on the social Semantic Web. First, we propose a two-step methodology for processing and analyzing social network data from the Semantic Web. Using the SPARQL query language, a derived RDF graph can be constructed that is tailored to a specific question. After a brief introduction to the notion of unevenness, this methodology is applied to examine unevenness in network properties of semantic data. Comparing Lorenz curves for different centrality measures, it is shown how examinations of unevenness can provide crucial hints regarding the topology of (social) Semantic Web data.

Key words: semantic Web, social network analysis, SPARQL, unevenness

1. Introduction. The *social Semantic Web* is a broad, non-technical term, referring to data on the Semantic Web (encoded in RDF) that contain social information. The most prevalent ontology on the social Semantic Web is the FOAF (Friend Of A Friend) vocabulary [9]. FOAF can express information "about people and the things they make and do" and especially about how they are related. In this article, we will use a socio-cultural ontology that is (partly) based on FOAF and also uses concepts from other well-known ontologies like Dublin Core.

The Semantic Web [5] in general is conceived as a large-scale distributed information system. While some constituents are still in development and its current uptake is relatively modest, the Semantic Web graph already shows the traits of a complex system. Complex systems are encountered in many different contexts and include such diverse examples as computer networks, social networks, neural networks and cellular networks [13]. As a complex system, the Semantic Web is characterized by [3, 17]:

- *Small world properties*: Made famous by Stanley Milgram's [25] letter experiment, the small world notion refers to the fact that the average shortest path length in a graph is very short (comparable to that of a random graph). In practice, this means that it takes only a few steps to reach any other (reachable) node in the network. It is advisable to also take the longest shortest path, known as the *diameter*, into account. During the last decade, several models have been proposed to account for the small-world effect [26, 31].
- *High clustering*: The neighbours of a given node are likely also neighbours of each other.
- *Skewed degree distribution*: The probability $P(k)$ that a node has degree k (is connected to k other nodes) is not randomly distributed. Instead, it follows a power law $P(k) \approx Ak^{-\gamma}$. Moreover, complex systems typically exhibit power law distributions in more than one way. With regard to the Semantic Web, previous research has shown that a diversity of relations—such as the relation between websites (domain names) and their number of Semantic Web documents or the relation between an ontology and its frequency of use—follows a power law [15].

These properties, however, raise several questions as well. In this article, we first discuss a two-step methodology for extracting the Semantic Web data (or 'semantic data' for short) that we are interested in from the rest. We then focus on the last characteristic and try to compare the skewedness of several network measures. We try to provide an answer to the following two research questions.

First, how can data on the social Semantic Web be used for Social Network Analysis (SNA)? Significant research in this area has already been performed by, among others, Ding et al. [15] and Peter Mika [23, 24]. Much work has concentrated on acquiring and aggregating data (often FOAF data),—especially merging information about unique persons turns out to be far from trivial. In the present article, we assume that 'clean' semantic data are already available and concentrate on the following step: the development of a methodology for using one single RDF graph as the 'master', which can be used as the basis for several kinds of SNA. Ideally, we want to keep as much information as possible and extract a multitude of potentially interesting relations. This particular aspect has received less attention so far.

Second, it is very rarely examined how skewed a distribution is. How can this notion be measured? Quantification of unevenness is crucial for a thorough understanding of a power law distribution; moreover, it can be used for comparison purposes between distributions and between networks.

*University of Antwerp, Informatie—en Bibliotheekwetenschap, CST, Venusstraat 35, 2000 Antwerpen, Belgium, raf.guns@ua.ac.be

Both questions will be discussed and demonstrated using semantic data from Agrippa. Agrippa is the catalogue and database of the Archive and Museum of Flemish Cultural Life (AMVC Letterenhuis, located in Antwerp, Belgium). Where applicable, the RDF version builds upon existing ontologies like FOAF and Dublin Core. Agrippa contains a wealth of information about both the archived materials and the socio-cultural actors (people and organizations) that have created them. We will mostly use Agrippa information about the 237,062 letters present at the AMVC Letterenhuis and their writers and recipients.

2. Two-step methodology. Semantic data can be stored in many different ways: as a (set of) document(s) in one of the many RDF syntaxes [4]; in a 'classic' relational database; or in a triplestore, a dedicated RDF database. For performance and convenience reasons, we are using a triplestore, but most techniques can also be performed on, for instance, RDF documents. The triplestore used is Sesame, freely available at `http://www.openrdf.org/`.[1]

Partly due to their distributed nature, semantic data may appear quite dazzling: many different kinds of data, drawn from several ontologies, between which a multitude of relations exist. How can one make heads or tails out of them? Assuming the existence of a set of fairly clearly defined questions to be answered, we propose a two-step methodology, which critically depends on the SPARQL query language [27] or a query language with similar capabilities. In short, the two steps are:

1. Construct an extraction query in SPARQL and apply it to the RDF graph. This yields a derived graph, specifically tailored to the question(s).
2. Convert the derived graph to a format intended for SNA.

We will now discuss both steps in greater detail, using a part of Agrippa as an example (shown in Figure 2.1). Both Organization and Person are a kind of Agent. A LetterContext ties together the different participants in the act of letter-writing: the writer(s), the recipient(s) and the letter as a physical object. A letter can be written and received by either an Agent or an AffiliationContext. This refers to a person (the 'affiliatee') acting on behalf of his/her affiliation to an organization (the 'affiliator').

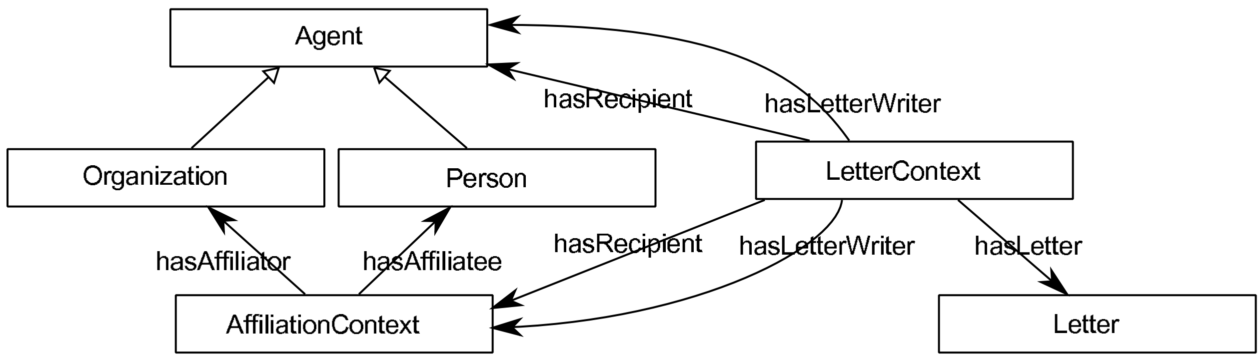

FIG. 2.1. *Part of the Agrippa ontology, showing the relations between six classes*

2.1. SPARQL information extraction. Four SPARQL query types exist: `SELECT`, `CONSTRUCT`, `ASK` and `DESCRIBE`. SPARQL queries are usually `SELECT` queries, which return a table of results. In this step, we employ `CONSTRUCT` queries, which return a new RDF graph. A similar architecture can also be found in the MESUR project [8, 28]. We will refer to the original graph as *source graph* and to the newly constructed graph as *derived graph*.

First, we compare the original graph in the triplestore and the questions to be answered. Some questions simply involve the extraction of parts of the RDF graph (ignoring the rest), like the following example. Suppose we want to examine only those letters that were created in an organizational context. This boils down to extracting the letters that are written by an Organization or an AffiliationContext:

```
PREFIX : <http://anet.ua.ac.be/agrippa#>
CONSTRUCT {
  ?context a                 :LetterContext ;
```

[1] For an overview of triplestores, see [20].

```
            :hasLetterWriter ?writer ;
            :hasRecipient    ?recipient ;
            :hasLetter       ?letter .
}
WHERE {
  ?context a                :LetterContext ;
            :hasLetterWriter ?writer ;
            :hasRecipient    ?recipient ;
            :hasLetter       ?letter .
  { ?writer a :Organization } UNION
  { ?writer a :AffiliationContext }
}
```

Other questions also require knowledge on how relations in the model interact,—these involve both extraction and combination of parts of the model. Here are two examples from Agrippa. The following query constructs a derived graph of persons and their affiliations to organizations. The result is a bipartite graph, i. e. a graph with two kinds of nodes (persons and organizations).

```
PREFIX : <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?person :affiliatedWith ?org }
WHERE {
  ?aff :hasAffiliator ?org ;
       :hasAffiliatee ?person .
}
```

And the following query constructs a simple derived graph that links author(s) and recipient(s) of each letter:

```
PREFIX : <http://anet.ua.ac.be/agrippa#>
CONSTRUCT { ?sender <urn:agrext#writesLetterTo> ?recipient }
WHERE {
  ?context :hasLetterWriter ?sender ;
            :hasRecipient    ?recipient .
}
```

It should be noted that it is often easier to obtain the desired results using one or more intermediate extraction queries. As such, a derived graph may become the source graph in a next step and so on. One could, for example, use the result of the first example as the source graph for the third example query. Although extraction queries are obviously not as powerful as a dedicated program or full-fledged reasoner, they are often sufficient and much faster to implement.

One of the advantages of storage in a triplestore is availability of the SPARQL protocol [14]. As its name implies, the SPARQL protocol is designed for exchanging SPARQL queries and results between clients and servers. It is entirely based on Web standards like HTTP and XML.

2.2. Conversion for SNA analysis. Once a derived graph has been obtained, it can be studied. There exist several projects for visualizing and exploring RDF and FOAF data, such as FOAF Explorer,[2] RDF-Gravity[3] and Visual Browser.[4] These tools, however, generally do not provide SNA measures like centrality and clustering, although Flink [23] seems a promising exception. Moreover, they generally do not scale to very large graphs. As long as there exist virtually no applications that successfully bring network analysis to RDF, it seems advisable to convert the derived graph to a more generic file format for network analysis.

Thus, while not strictly necessary, this step ensures compatibility with other SNA efforts and permits techniques that are difficult to perform on plain RDF graphs. We handle these conversions by integrating with pyNetConv, a Python library that can convert to most common formats, including Pajek, NetworkX, and GML.

[2]http://xml.mfd-consult.dk/foaf/explorer/
[3]http://semweb.salzburgresearch.at/apps/rdf-gravity/
[4]http://nlp.fi.muni.cz/projekty/visualbrowser/

3. Unevenness.

3.1. The Lorenz curve and the Gini evenness index. The distribution of degrees on the Semantic Web is—like many other relations—highly uneven: a small number of nodes has a huge amount of links, while the vast majority has very few. How can this unevenness be quantified?

Unevenness or inequality has been studied extensively in econometrics and informetrics. Since not all existing measures satisfy all necessary requirements [1, 16], we will limit the present discussion to two methods, using the following simple array as an example: $X = (1, 3, 4, 7, 10, 15)$. These numbers could express the distribution of wealth, the number of publications per author or the number of links per node. Clearly, there is some unevenness, but how much exactly?

The Lorenz curve [21] is a graphical representation of unevenness. First, we determine the relative amounts:

$$a_i = \frac{x_i}{\sum x}$$

resulting in (1/40, 3/40, 1/10, 7/40, 1/4, 3/8). The horizontal axis of the Lorenz curve has the points i/N ($i = 1, 2, \ldots, N$). The vertical axis of the Lorenz curve has their cumulative fraction: $a_1 + a_2 + \ldots + a_i$. We thus construct the Lorenz curve (Figure 3.1). The diagonal line represents the case of perfect evenness—everyone possesses the same amount. The further the curve is removed from the diagonal, the greater the unevenness. Note that we have ranked our numbers in increasing order, resulting in a convex Lorenz curve. The concave Lorenz curve results from ranking in decreasing order and is completely equivalent. Complete unevenness—one person has everything, and the rest nothing—would be represented as a convex curve following the bottom and the right side of the plot.

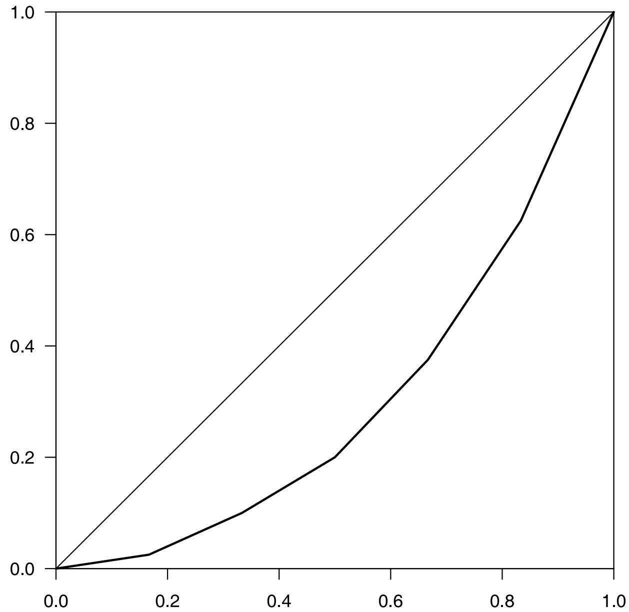

FIG. 3.1. *Convex Lorenz curve of the array (1, 3, 4, 7, 10, 15)*

Suppose we want to express this unevenness in a number. A good measure is the Gini evenness index G' [29], originally devised to characterize the distribution of wealth over social classes [18],

$$G'(X) = \frac{2}{\mu N^2} \left(\sum_{j=1}^{N} (N + 1 - j) x_j \right) - \frac{1}{N}$$

with x_j ranked in increasing order and μ the mean of the set x_j. There exists a direct relation between the Lorenz curve and the Gini evenness index: G' is equal to twice the area under the convex Lorenz curve.

Lorenz curves determine a partial order: in some, but not all, cases, an order can be determined from the comparison of two Lorenz curves. Indeed, if one convex Lorenz curve is completely below another, then the former expresses less evenness than the latter. It should be stressed that Lorenz curves may 'overlap' or cross each other. In these cases, no order can be determined from the curves [29].

3.2. Application to Agrippa.

3.2.1. Overview of network measures. Let us take the author-recipient graph constructed in the last example of 2.1 $N = 40,914$ as an example. Each node is connected by 5.08 links on average, but the actual in- and out-degree follow a power law distribution (Figure 3.2). We will consider the following network measures, most of which are defined by Wasserman & Faust [30]:

- *Degree centrality (DC)*: is the number of links connected to a given node.
- *Betweenness centrality (BTC)*: characterizes the importance of a given node for establishing short pathways between other nodes.
- *Closeness centrality (CC)*: characterizes how fast other nodes can be reached from a given node.
- *Pagerank (PR)*: characterizes the importance of a given node by combining its number of in-links with the importance of the nodes that link to it. The algorithm was originally created for determining a web page's importance [10] but has since been used in many other contexts as well (e.g., [12, 22]).

This small list of measures is in no way intended to be exhaustive. Many other measures exist and even the ones listed here have several varieties themselves. They have been chosen because they are both well-known and generally used and accepted. Moreover, they can be computed using standard software tools. For the current article, we used the *igraph* R package, available at `http://cneurocvs.rmki.kfki.hu/igraph/`.

The centrality measures listed above all have variants for directed and undirected networks, but we will only consider the directed variants. Both degree centrality and closeness centrality have different algorithms for in-links and out-links. We can distinguish between in-degree centrality (IDC) and out-degree centrality (ODC), and between in-closeness centrality (ICC) and out-closeness centrality (OCC). This distinction is not useful for betweenness centrality and PageRank.

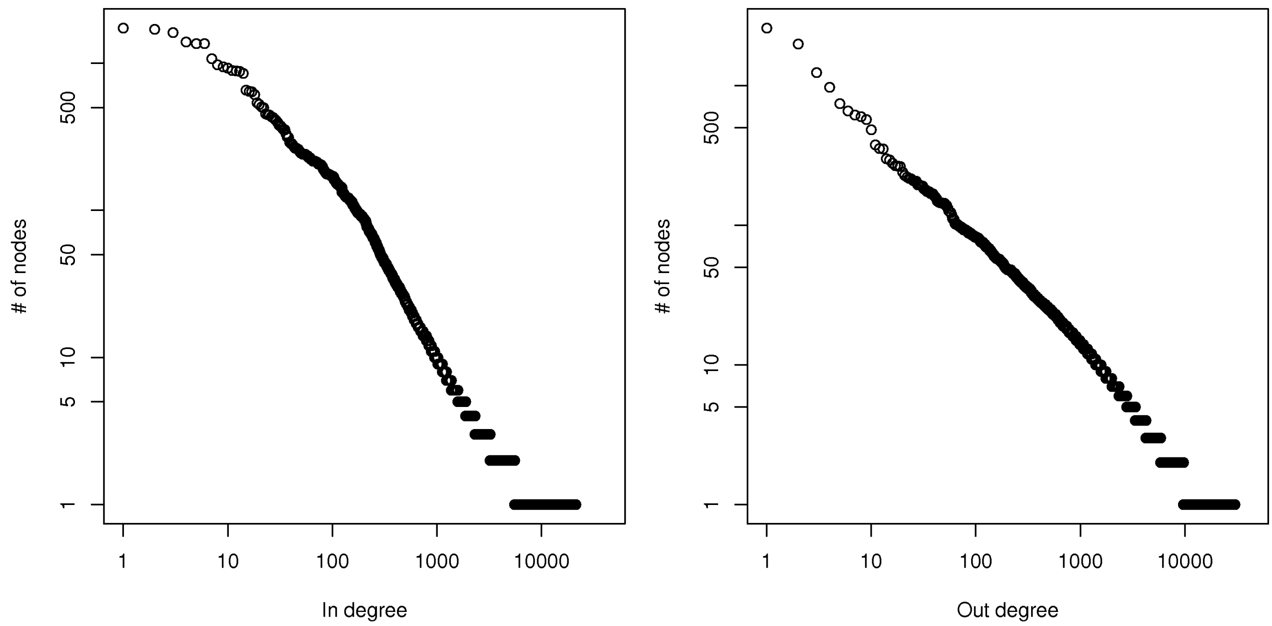

FIG. 3.2. *Power law distribution for in-degree and out-degree*

3.2.2. Comparison of unevenness between network measures. The graph is not fully connected, but the main component ($N = 40,303$) accounts for the vast majority of nodes (98.5%). Henceforth, we will only consider the nodes that are part of the main component, since very small components (e.g., $N = 2$) can distort the overall picture. For instance, a node v in such a component may have $CC_v = 1$, even if its position in the overall network is obviously marginal. We therefore consider it methodologically more correct to only consider nodes that are part of the main component.

Comparing IDC to ODC and ICC to OCC (Figure 3.3), we see that in both cases the measure based on in-links is more uneven. In spite of this difference, it should be noted that in both cases the shape of the Lorenz curve of the in-link-based measure is similar to that of the out-link-based one.

PageRank is, in a sense, a more refined version of in-degree centrality. Whereas the latter only considers the local neighbourhood (i. e. the number of links to a given node), PageRank also considers the status of the nodes that are linking to a given node by iteratively passing status between nodes. Figure 3.4 shows that PageRank is actually more even than in-degree centrality. In other words: some extreme variations in degree are 'evened out' by looking at a node's status in the entire network rather than just its number of in-links. Inspection of the data reveals that this is almost exclusively due to nodes with a low number of in-links from some very high status nodes. Put another way, differences between PageRank and IDC may be due to IDC either 'overrating' or 'underrating' some nodes; at least for this example, the latter is mostly the case. Despite the outliers, PageRank and in-degree centrality are highly correlated. Figure 3.4 also illustrates the usefulness of the Lorenz curve for comparing different measures: it makes it possible to, for instance, compare raw numbers (IDC) to normalized ones (PageRank).

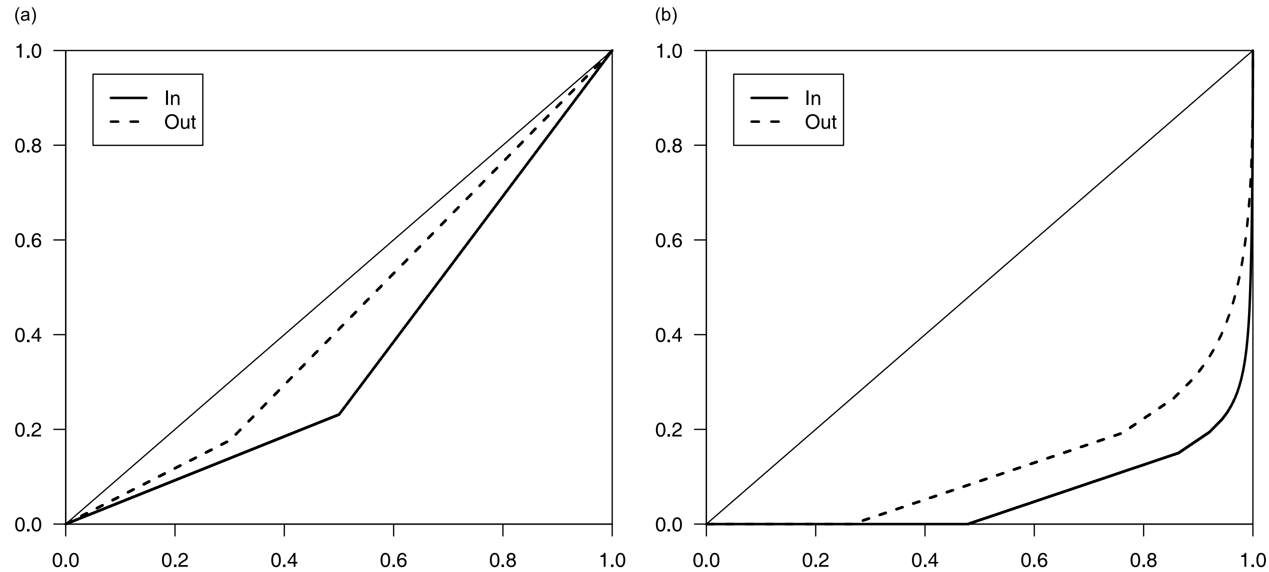

Fig. 3.3. *Comparison of unevenness between in-link-based and out-link-based measures. (a) Comparison of ICC to OCC, (b) Comparison of IDC to ODC*

Betweenness centrality is remarkably uneven (Figure 3.5). Indeed, we immediately see that more than 80% of all nodes have zero betweenness centrality. The Lorenz curve clearly reveals that betweenness centrality is considerably less even than any of the other measures discussed here.

3.3. Discussion. Comparing the Lorenz curves of the different centrality measures reveals a remarkably diversified picture. Betweenness centrality is clearly least even of all. Subsequently, we get degree centrality, PageRank and closeness centrality. The Gini evenness indices basically tell the same story and are summarized in Table 3.1.

As a tentative explanation, we suggest that these differences may be largely due to the small-world effect [26, 31]. Even marginal nodes are relatively close to all others, accounting for minimal differences in closeness. Indeed, the length of the diameter—the longest shortest path—is only 11 and the average shortest path length only 4.12!

As a whole, the graph fits well into the bow-tie or corona models [6, 7, 11], which were originally devised for modelling and explaining link structure on the World Wide Web. The core of the main component is the Largest Strongly Connected Component or LSCC ($N = 9,723$), a component in which any node can be reached (obeying the direction of the links). The LSCC itself has a nucleus of hubs [13, 19], through which almost all other shortest paths pass. These hub nodes typically have extremely high degree centrality. This has two

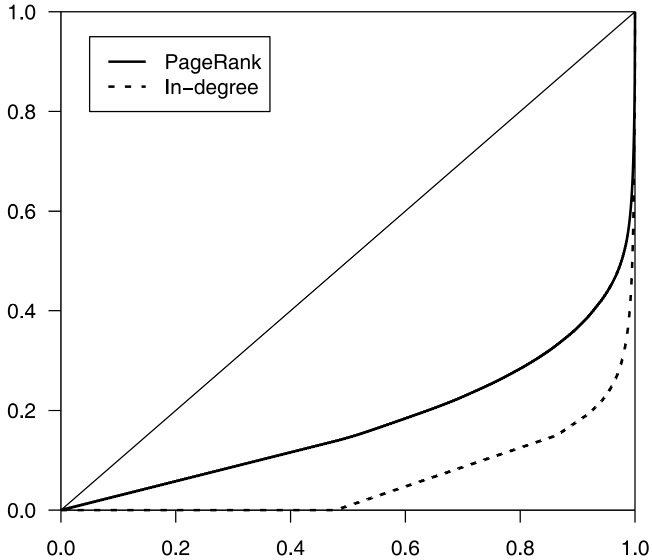

FIG. 3.4. *Comparison of unevenness between PageRank and in-degree centrality*

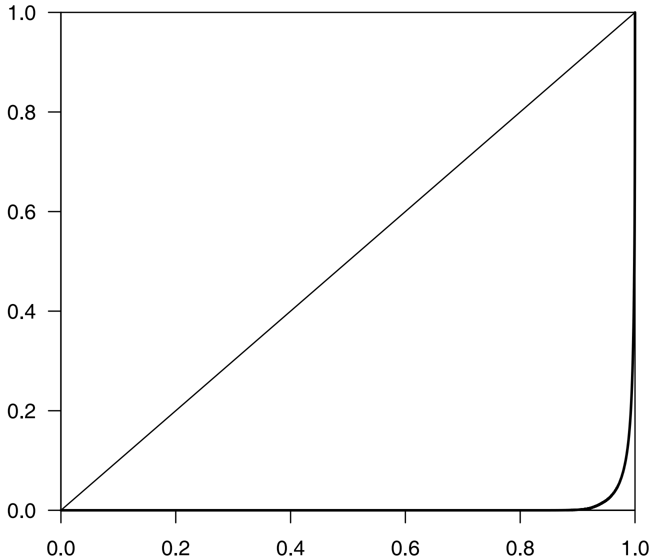

FIG. 3.5. *Unevenness of betweenness centrality*

interesting, seemingly opposite, effects. On the one hand, closeness is increased and closeness centrality becomes more even. On the other hand, it brings about a very uneven betweenness centrality distribution.

PageRank distribution is more even than one might intuitively expect. The hubs have a high status, which is partially transmitted to each of the nodes they link to. As such, a large number of nodes gains a higher PageRank than might be expected from their in-degree centrality or betweenness centrality. Indeed, even if no shortest paths pass through them, their PageRank will still be relatively high. This property of PageRank is very desirable for ranking Web pages, but may be unwanted in some applications of SNA.

4. Conclusions. We have shown how SPARQL can be used in processing social Semantic Web data in a simple two-step methodology, converting the source graph to a better suited derived graph. While SPARQL is obviously less powerful than a 'real' reasoning engine or a dedicated program, it is often sufficient and may well prove simpler and faster to implement. RDF tools are generally not geared towards SNA, although Flink [23] incorporates some basic SNA statistics. Therefore, conversion to other formats is currently recommendable but, luckily, straightforward.

Table 3.1
Gini evenness index of all centrality measures in increasing order of evenness

Centrality measure	G'
Betweenness centrality	0.01
In-degree centrality	0.12
Degree centrality (in and out)	0.25
Out-degree centrality	0.26
PageRank	0.35
In-closeness centrality	0.73
Out-closeness centrality	0.88
Closeness centrality (in and out)	0.94

The Lorenz curve and the Gini evenness index G' are two excellent methods for studying unevenness. Taking Agrippa as a concrete example, it can be seen that unevenness measures may confirm or enforce hypotheses regarding the network topology. In the example discussed, the massive difference between betweenness centrality and closeness centrality distribution confirms the small-world hypothesis and reveals the topology of the graph with a small nucleus, through which most other paths must pass. The example also illustrates the need for a wide variety of centrality measures: they are indeed very different (as is obvious from just comparing the Lorenz curves) and each reveals a different aspect of the network.

Most of these results, such as the establishment of the small-world effect, could have been achieved without studying the unevenness of network properties. Consequently, the current paper should be regarded as a first step: it illustrates how unevenness measures can be used to achieve results similar to existing, well-established methods. In future research, we hope to expand upon these results by studying a greater variety of (social) networks, including different classes of small-world networks [2].

Acknowledgements. A previous version of this paper was presented at the Social Aspects of the Web 2008 Workshop, held in conjunction with the BIS2008 conference in Innsbruck. I thank prof. Ronald Rousseau and the organizers and participants to the Workshop for their constructive comments, and prof. Richard Philips for providing access to the Agrippa dataset.

REFERENCES

[1] P. D. Allison, *Measures of inequality*, American Sociological Review, 43 (1978), pp. 865–880.

[2] L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley, *Classes of small-world networks*, Proceedings of the National Academy of Sciences, 97 (2000), pp. 11149–11152.

[3] D. Bachlechner and T. Strang, *Is the semantic web a small world?*, in Proceedings of the Second International Conference on Internet Technologies and Applications (ITA 07), 09 2007.

[4] D. Becket, *New syntaxes for rdf.* `http://www.dajobe.org/2003/11/new-syntaxes-rdf/paper.html` 2004.

[5] T. Berners-Lee, J. Hendler, and O. Lassila, *The Semantic Web*, Scientific American, 284 (2001), pp. 28–37.

[6] L. Björneborn, *Small-World Link Structures across an Academic Web Space: A Library and Information Science Approach*, PhD thesis, Royal School of Library and Information Science, Denmark, 2004.

[7] ———, *'mini small worlds' of shortest link paths crossing domain boundaries in an academic web space*, Scientometrics, 68 (2006), pp. 395–414.

[8] J. Bollen, M. A. Rodriguez, H. Van de Sompel, L. Balakireva, W. Zhao, and A. Hagberg, *The largest scholarly semantic network... ever.*, in Proceedings of the 16th international conference on World Wide Web, ACM Press New York, NY, USA, 2007, pp. 1247–1248.

[9] D. Brickley and L. Miller, *Foaf vocabulary specification 0.91. namespace document 2 november 2007 Ũ openid edition.* http://xmlns.com/foaf/spec/, 2007.

[10] S. Brin and L. Page, *The anatomy of a large-scale hypertextual web search engine*, Computer Networks and ISDN Systems, 30 (1998), pp. 107–117.

[11] A. Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins, and J. Wiener, *Graph structure in the web*, Computer Networks, 33 (2000), pp. 309–320.

[12] P. Chen, H. Xie, S. Maslov, and S. Redner, *Finding scientific gems with google's pagerank algorithm*, Journal of Informetrics, 1 (2007), pp. 8–15.

[13] C. Christensen and R. Albert, *Using graph concepts to understand the organization of complex systems*, Nov 2006.

[14] K. G. Clark, L. Feigenbaum, and E. Torres, *Sparql protocol for rdf w3c recommendation 15 january 2008.* `http://www.w3.org/TR/rdf-sparql-protocol/` 2008.

[15] L. Ding and T. Finin, *Characterizing the semantic web on the web*, in Proceedings of the 5th International Semantic Web Conference, 2006.

[16] L. Egghe and R. Rousseau, *Transfer principles and a classification of concentration measures*, Journal of the American Society for Information Science, 42 (1991), pp. 479–489.
[17] R. Gil and R. García, *Measuring the semantic web*, in Advances in Metadata Research, Proceedings of MTSR'05, 2006.
[18] C. Gini, *Il diverso accrescimento delle classi sociali e la concentrazione della richezza*, Giornale degli Economisti, 11 (1909).
[19] J. M. Kleinberg, *Authoritative sources in a hyperlinked environment*, in SODA '98: Proceedings of the ninth annual ACM-SIAM symposium on Discrete algorithms, Society for Industrial and Applied Mathematics, 1998, pp. 668–677.
[20] R. Lee, *Scalability report on triple store applications.* `http://simile.mit.edu/reports/stores/` 2004.
[21] M. O. Lorenz, *Methods of measuring the concentration of wealth*, Publications of the American Statistical Association, 9 (1905), pp. 209–219.
[22] R. Mihalcea, P. Tarau, and E. Figa, *Pagerank on semantic networks, with application to word sense disambiguation*, in COLING '04: Proceedings of the 20th international conference on Computational Linguistics, Morristown, NJ, USA, 2004, Association for Computational Linguistics, p. 1126.
[23] P. Mika, *Flink: Semantic Web technology for the extraction and analysis of social networks*, Journal of Web Semantics, 3 (2005), pp. 211–223.
[24] P. Mika, T. Elfring, and P. Groenewegen, *Application of semantic technology for social network analysis in the sciences*, Scientometrics, 68 (2006), pp. 3–27.
[25] S. Milgram, *The small world problem*, Psychology Today, 2 (1967), pp. 60–67.
[26] M. E. J. Newman, *Models of the small world*, Journal of Statistical Physics, 101 (2000), pp. 819–841.
[27] E. Prud'hommeaux and A. Seaborne, *Sparql query language for rdf. w3c recommendation 15 january 2008.* `http://www.w3.org/TR/rdf-sparql-query/` 2008.
[28] M. A. Rodriguez, J. Bollen, and H. Van de Sompel, *A practical ontology for the large-scale modeling of scholarly artifacts and their usage*, in JCDL '07: Proceedings of the 2007 Conference on Digital Libraries, New York, NY, USA, 2007, ACM Press, pp. 278–287.
[29] R. Rousseau, *Lorenz curves determine partial orders for comparing network structures*, in CREEN (Critical events in evolving networks) Workshop, 2007.
[30] S. Wasserman and K. Faust, *Social Network Analysis: Methods and Applications*, Structural Analysis in the Social Sciences, Cambridge University Press, 1994.
[31] D. J. Watts and S. H. Strogatz, *Collective dynamics of 'small-world' networks.*, Nature, 393 (1998), pp. 440–442.

Edited by: Dominik Flejter, Tomasz Kaczmarek, Marek Kowalkiewicz
Received: February 2nd, 2008
Accepted: March 20th, 2008
Extended version received: July 7th, 2008