A Structural Holistic Believability Metric for Influence in Linked Data

TREO Talk Paper

Carolyn C. Matheus
Marist College
Carolyn.Matheus@Marist.edu

Alan G. Labouseur
Marist College
Alan.Labouseur@Marist.edu

Abstract

We live in an evolving world of rapidly accumulating and highly granular big data, constantly growing in volume and increasing in velocity. This TREO paper focuses on the data quality dimension *believability*, specifically as it applies to evaluating influence in linked data. Believability is highly relevant in social media and other forms of linked data. Referring to the extent to which data are regarded as credible, believability is closely related to the fourth "V" of big data, *veracity*, which describes accuracy and trustworthiness (Shankaranarayanan and Blake, 2017), and contributes to the fifth "V" *value*. Prior work has focused on *provenance*-based (Prat and Madnick, 2008), *context*-based (Serra and Marotta, 2016), and *reputation*-based (Cai and Zhu, 2016) approaches to believability. In each case, their efficacy is situational, depending on the specific data under analysis. We propose a **structural**-based approach, exploiting the fact that regardless of its dynamic content and meta-content (e.g., provenance, context, reputation), the structure of linked data remains the same (if not, it ceases to be linked data). We illustrate our structural approach to believability by analyzing influence in linked data using a network from Yelp, an online linked directory service and crowd-sourced review forum largely about food.

There are many ways to understand and measure influence in linked data. Consider one particular person in a graph containing people. We might be interested in determining their influence by looking at the number of their immediate friends (which can be structurally calculated by determining their vertex degree), along with how well connected and relevant those friends are (which can be structurally calculated through clustering coefficient and PageRank, respectively). Each of those metrics provide insight into influence (or lack thereof) within linked data, but none present the whole picture because there are ways to artificially inflate or otherwise "game" those individual measures. However, a **holistic** approach for evaluating the believability of influence measures in linked data seems attainable by combining individual measures. We will present phase one of this research-in-progress: combining graph analytics to develop a *Structural Holistic Believability Metric* for influence in linked data.

Intuitively, incorporating multiple believability measures seems like it could increase data quality by improving credibility and value judgements of influence in linked data. But how do we **know** that? How can we test and validate it? Generating a sense of believability is difficult, as it is an inherently human concept. Although this human ability seems like it might be helpful for validating believability metrics, it is not, because the volume, velocity, and variety inherent in big data is too great for humans to process. We will encourage discussion of this important topic, which addresses phase two of this research-in-progress.

References

Cai, Y., & Zhu, D. (2016). "Reputation in an open source software community: Antecedents and impacts" *Decision Support Systems*, vol. 91, pp. 103-112.

Prat, N., & Madnick, S. (2008). "Measuring data believability: A provenance approach" P*roceedings of the 41st Annual Hawaii International Conference on System Sciences*, pp. 393.

Serra, F., & Marotta, A. (2016). "Data Warehouse Quality Assessment Using Contexts" *Proceedings of the 17th International Conference on Web Information Systems Engineering,* pp. 436-448.

Shankaranarayanan, G., & Blake, R. (2017). "From Content to Context: The Evolution and Growth of Data Quality Research" *Journal of Data and Information Quality* (8:2), pp. 9:1-9:28.

A Structural Holistic Believability Metric for Influence in Linked Data

AMCIS 2020 TREO

A Structural Holistic Believability Metric for Influence in Linked Data

TREO Talk Paper

Carolyn C. Matheus
Marist College
Carolyn.Matheus@Marist.edu

Alan G. Labouseur
Marist College
Alan.Labouseur@Marist.edu

Abstract

We live in an evolving world of rapidly accumulating and highly granular big data, constantly growing in volume and increasing in velocity. This TREO paper focuses on the data quality dimension *believability*, specifically as it applies to evaluating influence in linked data. Believability is highly relevant in social media and other forms of linked data. Referring to the extent to which data are regarded as credible, believability is closely related to the fourth "V" of big data, *veracity*, which describes accuracy and trustworthiness (Shankaranarayanan and Blake, 2017), and contributes to the fifth "V" *value*. Prior work has focused on *provenance*-based (Prat and Madnick, 2008), *context*-based (Serra and Marotta, 2016), and *reputation*-based (Cai and Zhu, 2016) approaches to believability. In each case, their efficacy is situational, depending on the specific data under analysis. We propose a **structural**-based approach, exploiting the fact that regardless of its dynamic content and meta-content (e.g., provenance, context, reputation), the structure of linked data remains the same. (If not, it ceases to be linked data.) We illustrate our structural approach to believability by analyzing influence in linked data using a network from Yelp, an online linked directory service and crowd-sourced review forum largely about food.

There are many ways to understand and measure influence in linked data. Consider one particular person in a graph containing people. We might be interested in determining their influence by looking at the number of their immediate friends (which can be structurally calculated by determining their vertex degree), along with how well connected and relevant those friends are (which can be structurally calculated through clustering coefficient and PageRank, respectively). Each of those metrics provide insight into influence (or lack thereof) within linked data, but none present the whole picture because there are ways to artificially inflate or otherwise "game" those individual measures. However, a **holistic** approach for evaluating the believability of influence measures in linked data seems attainable by combining individual measures. We will present phase one of this research-in-progress: combining graph analytics to develop a *Structural Holistic Believability Metric* for influence in linked data.

Intuitively, incorporating multiple believability measures seems like it could increase data quality by improving credibility and value judgements of influence in linked data. But how do we **know** that? How can we test and validate it? Generating a sense of believability is difficult, as it is an inherently human concept. Although this human ability seems like it might be helpful for validating believability metrics, it is not, because the volume, velocity, and variety inherent in big data is too great for humans to process. We will encourage discussion of this important topic, which addresses phase two of this research-in-progress.

References

Cai, Y., & Zhu, D. (2016). "Reputation in an open source software community: Antecedents and impacts" *Decision Support Systems*, vol. 91, pp. 103-112.

Prat, N., & Madnick, S. (2008). "Measuring data believability: A provenance approach" *Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pp. 393.

Serra, F., & Marotta, A. (2016). "Data Warehouse Quality Assessment Using Contexts" *Proceedings of the 17th International Conference on Web Information Systems Engineering*, pp. 436-448.

Shankaranarayanan, G., & Blake, R. (2017). "From Content to Context: The Evolution and Growth of Data Quality Research" *Journal of Data and Information Quality* (8:2), pp. 9:1-9:28.

Carolyn C. Matheus
Carolyn.Matheus@Marist.edu

Alan G. Labouseur
Alan.Labouseur@Marist.edu

Prior work relies on situational efficacy:
- *provenance*-based
- *context*-based
- *reputation*-based

We propose a **structural** approach, exploiting the fact that regardless of its dynamic content and meta-content the structure of linked data remains the same.

!

Transforming data from Yelp using JavaScript, we created linked data in the form of graphs for G*.

{JSON} <JavaScript/> G* The Dynamic Graph Database

With the data now fit for use, we executed graph queries to compute influence for each person by looking at the number of their immediate friends (vertex degree) along with how well connected and relevant those friends are (clustering coefficient and PageRank).

```
select vertex.id as "vid",
       vertex.total_degree
from "default"
where graph.id = 1
order by "vid" asc
```

compiles to

```
operator vertex@*=VertexOperator([],[1])
operator projection@*=ProjectionOperator([vertex@local],
   [cardinality(outgoing_edges)+cardinality(incoming_edges)],
   [vid,total_degree])
operator union@0=UnionOperator([projection@*])
operator sort@0=SortOperator([union@0], [vid:asc])
run sort@0
```

Ranking the top-10 anonymous user ids by ...
(a) total degree
(b) clustering coefficient
(c) PageRank
... we see little overlap among them. No one measure captures holistic influence. But combining them in some way seems promising.

WmAyExqSWoiYZ5XEqpk_Uw
AaZdXn0I6F5bdIVwGpxdDA
nKoB5cWZHXYUIUcQsUDogA
spJUPXI7QaIctU0FO5c42w
^ANkfLbDf8aiBQ7vywIL6w
neOOSMNWcVL0o2Xwb0goVg
fczQCSmaWF78toLEmb0Zsw
0IAOkW3KD1Dsx2hnwb0CSA
GJrGPKF2xxB06Es6aH1VWg
kGgAARL2UmvCcTRfiscjug
Total Degree
Number of Friends

(a)

01GNfaGbQ5iXXGQqsMNT5A
46SGCG3pkXq^eEucAOCT_g
1^t3fFZp_HaM^yj57sTbpQ
9lk_zFB8UtJrX1^vbRzFMQ
Ar7G22UaiKIv^c^pu2t_OQ
3P9Y7hKlLzbXjXn_vPqHgw
icP5H8hsXfhz5e9tPojVIQ
1MMYXVCUDuC5wGglauM^Kw
JZqLeJW^rVqwHpd4qGV6HA
285hxd_9FUmmO1pmJGEsAQ
Clustering Coefficient
Connectivity

(b)

zZlHqWiCrCj0WKSNI1Nxlw
ZzpzwgRKp7MeuNKghd54NQ
zZxr7X10CDThXZbnkLmNVA
ZZ43etAB2n_T53YBYtf8Dw
ZkuPKOz3tN9iTUZrGzj3nQ
PWj9W9lYnSXazgafWejyQ
WmAyExqSWoiYZ5XEqpk_Uw
zfb_dSwWV5mV4f_ZAgkYbg
zTWH9b_ItSdLOK9ypeFOIw
ZPolhetd6Od5_VhXPbFIxw
PageRank
Relevancy

(c)

Combining structural individual measures to form a holistic believability metric seems like a good idea.

But how can we know?

Combining measures is easy. Validating the metric is hard. Believability is a human concept. But the volume, velocity, and variety inherent in big data makes it impossible for humans to judge it all.

?

Discussion: How might we validate this approach?