# Connecting the Dots: Explaining Relationships Between Unconnected Entities in a Knowledge Graph

Nitish Aggarwal*, Sumit Bhatia°, and Vinith Misra°

*Insight Centre for Data Analytics, National University of Ireland, Galway, Ireland
°IBM Research, Almaden, USA
nitish.aggarwal@insight-center.org,{sumit.bhatia,vmisra}@us.ibm.com

**Abstract.** We discuss the problem of explaining relationships between two unconnected entities in a knowledge graph. We frame it as a path ranking problem and propose a path ranking mechanism that utilizes features such as specificity, connectivity, and path cohesiveness. We also report results of a preliminary user evaluation and discuss a few example results.

## 1 Introduction

The advent of semantic knowledge bases like DBpedia, Freebase, etc. has led to the development of smart search systems that produce rich and enhanced results by providing additional related information about the entities/concepts being queried by the users. Further, increasing efforts are being made to build *knowledge discovery systems* that help users to navigate/explore the semantic graph and discover hitherto unknown, yet extremely useful information. For example, Nagarajan et al. [6] describe a discovery system that uses a semantic network built out of medical literature and helps researchers in discovering previously unknown protein-protein interactions. Likewise, web search engines like Google, Bing, etc. also incorporate data from their knowledge graphs to provide a list of entities that are related to the user search query [3] and users often navigate through these recommended entities to discover new non-trivial information about their search topics.

Despite providing a greatly simplified knowledge discovery process by recommending related entities of interest, such systems often fail to provide explanations for such recommendations to users, especially for less popular entities and entities that are not directly connected to the input entity. For example, for an entity query "Abu Bakr al-Baghdadi"(leader of the terrorist organization Islamic State of Iraq and the Levant (ISIL)), Google recommends entities such as "Musab al-Zarqawi", "Qasem Soleimani", etc., but fails to provide any explanation about how these entities are related to the input entity. Previous research efforts [5, 7] have tried to explain the relatedness between entities by deriving important paths between entities in the knowledge graph. However, these methods generally focus either on popular entities in the graph or rely on query log data from the search engines that may not be available always, especially in enterprise domains.

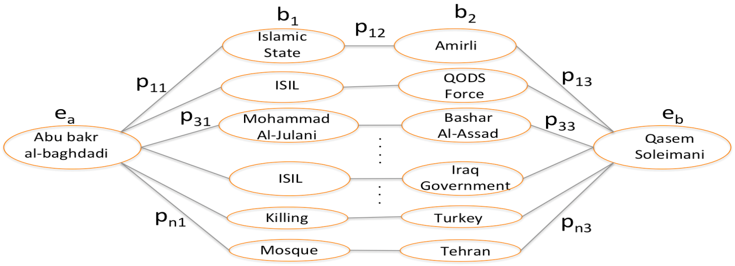

**Fig. 1.** Paths between “Abu bakr al-baghdadi” and “Qasem Soleimani”

In this work, we address the problem of *ranking all the paths between any two entities in a knowledge graph*. A solution to this problem can help in explaining relationships between seemingly unconnected entities as well as in finding interesting, non-obvious paths between two given entities. Given two entities in the graph, there could be hundreds and thousands of possible paths connecting the two entities. We posit that each such path represents a fact (or a hidden connection), and thus, provides a potential explanation of the relationship between the two entities. Not all the paths connecting the two entities are equally important and therefore, a mechanism is required to *rank* these paths based on their usefulness. For instance, figure 1 shows several different paths existing in our knowledge graph between entities “Abu bakr al-baghdadi” and “Qasem Soleimani” (a major general in the Iranian Army of the Guardians of the Islamic Revolution). We observe that all the paths may not be considered equally informative as some of the paths involve very generic entities like “Mosque”, “Tehran”, etc. and some paths involve very specific entities like “Islamic State” and “Amirli”, representing the fact that Islamic State and Iranian Army led by Qasem Soleimani were involved in a battle in the city of Amirli. Thus, we see how these paths can provide deep insights into different type of relations between given entities.

## 2 Proposed Path Ranking Algorithm

Let $e_a$ and $e_b$ be the two input entities and let $P = \{P_1, P_2, \ldots, P_n\}$ be the set of all possible paths connecting $e_a$ and $e_b$. Let $P_i = \{p_{i1}, p_{i2}, \ldots, p_{im}\}$ be a path of length $m$ between $e_a$ and $e_b$, consisting of $m-1$ bridging nodes $b_1, b_2 \ldots b_{m-1}$, such that $p_{i1}$ is the edge between $e_a$ and $b_1$, $p_{i2}$ is the path between $b_1$ and $b_2$, and so on. Our task is to produce a ranked list of paths in $P$ ordered by their relevance scores. Our proposed ranking function is based on the intuition that a relevant and informative path consists of relevant and useful bridging nodes and edges and we utilize following signals to capture this intuition in the relevance score of a given path between $e_a$ and $e_b$.

**Specificity:** This measure is analogous to the inverse document frequency (IDF) concept used frequently in information retrieval models. There are many popular entities that are connected to a disproportionately large number of other entities in the graph. For example, *USA* is connected to a large number of other entities corresponding to different countries, persons, organizations, etc. A path connected through such highly popular bridging nodes may not be as informative and useful as a path connected through relatively rare bridging nodes. Specificity of a given entity is computed as the inverse of total neighbor count of the given entity and specificity of a given path is, in turn, computed as sum of specificity scores of each bridging node in the path.

**Connectivity:** This measure tries to capture the contribution of constituent edges to the overall relevancy of a given path. A path consisting of stronger edges is more probable to be useful than a path consisting of weaker edges. We posit that the strength of an edge connecting two entities is directly proportional to *relatedness* of the two entities. Computing relatedness scores between two entities in knowledge graphs is a well addressed problem [1,3] and we use the distributional semantics model (DSM) [2] to compute the relatedness score of two connected entities and hence, their *edge strength*. We generate the DSM vector for each entity over Wikipedia concepts and compute the relevance scores between two entities by calculating cosine scores between their vectors.

**Cohesiveness:** While the previous two signals were concerned with the contribution of individual bridging nodes and constituent edges to the overall relevancy score of a path, this measure reflects the strength of *linkages* between adjacent edges in the path. Connectivity measure as discussed above provides the strength of individual edges in *isolation* and hence, may not capture the relevancy of different edges in context of others. This measure, therefore, tries to capture the *cohesiveness* of successive edges that form a given path as follows. For a pair of consecutive edges $p_i$ and $p_{i+1}$ connecting entities $e_a, e_b$, and $e_c$, respectively, we obtain the composite DSM vector of entity $e_a$ and $e_b$ by adding their individual DSM vectors and then take the cosine of the DSM vector $e_c$ with this composite vector. This way, cohesiveness score provides the relevancy of entity $e_c$ to $e_a$ in context of adjacent bridging entity $e_b$. The overall cohesiveness score of the path is then computed by summing over the cohesiveness score of each consecutive pair of edges in the path.

Finally, the overall relevance score of a given path is obtained by taking a product of all the above three scores.

## 3 Evaluation

We use a semantic graph constructed from text of all articles in Wikipedia by automatically extracting the entities and their relations by using IBM's Statistical Information and Relation Extraction (SIRE) toolkit[1]. Even though there exist popular knowledge bases like DBPedia that contain high quality data, we chose to construct a semantic graph using automated means as such a graph will be closer to many real world scenarios where domain specific data is used and high quality curated graphs are often not available. Our graph contains more than 30 millions entities and 192 million distinct statements in comparison to 4.5 million entities and 70 million statements in DBpedia.

Path ranking in knowledge graphs is a relatively new problem and datasets used in previous related research [4,7] for explaining relationships mainly focus on popular paths and ignore rare and inconspicuous relationships. Further, there may not be one single relevant path (or explanation) as the two entities could be related in multiple different ways. In this work, we perform a preliminary qualitative evaluation through a

[1] http://ibmlaser.mybluemix.net/siredemo.html

user study involving three human assessors. We asked each of the three volunteers to query our system with three entity pairs of their interest and showed them top 20 paths ranked by their relevance scores as described above. The evaluators were then asked to rate each path using the following criterion – (0) non-relevant path, (1) relevant path, somewhat informative (2) relevant and highly informative path. There are a total of 180 paths (9 pairs times 20 paths) to be evaluated and each evaluator provided judgments for 60 paths. The evaluators rated around 15% of the paths non-relevant, 75% as relevant and 10% as relevant paths with discovery. As an example of results, the top ranked path of length 3 between entities "Abu Bakr Al-Baghdadi" and "Qasem Soleimani" is connected through entities "Islamic States" and "Amirli" and corresponds to the fact (taken from Wikipedia) that *"...the fight over Amirli in eastern Iraq has been one of the most important battles against ISIS. The response to ISIS's push against the town was likely formulated by Qassem Suleimani, the head of the Iranian Revolutionary Guards Corps' Qods Force ..."* Another example of path illustrating the potential of proposed approach in unravelling interesting and hidden facts is the top ranked path between an Indian actor "Aamir Khan" and Hollywood director "Christopher Nolan" which reveals the fact that *"Aamir Khan's movie "Ghajini" was the remake of Hollywood movie "Momento" directed by "Christopher Nolan."*

## 4 Conclusion and Future Work

We presented a path ranking mechanism to explain relatedness of two unconnected entities in a knowledge graph and to uncover hidden connections between the two entities. We used specificity, connectivity and cohesiveness features to measure the quality of a path and performed a small scale, preliminary evaluation of the proposed approach. The results from this preliminary study are encouraging and provide some support to the proposed approach's ability to find high quality paths between entities. However, a more rigorous qualitative and quantitative evaluation is required to confirm these initial findings and this will be the focus of our future work.

## References

1. N. Aggarwal, K. Asooja, H. Ziad, and P. Buitelaar. Who are the american vegans related to brad pitt?: Exploring related entities. In *Proceedings of the 24th International Conference on World Wide Web Companion*, 2015.
2. N. Aggarwal and P. Buitelaar. Wikipedia-based distributional semantics for entity relatedness. In *2014 AAAI Fall Symposium Series*, 2014.
3. R. Blanco, B. B. Cambazoglu, P. Mika, and N. Torzec. Entity recommendations in web search. In *International Semantic Web Conference (2)*, pages 33–48, 2013.
4. L. Fang, A. D. Sarma, C. Yu, and P. Bohannon. Rex: explaining relationships between entity pairs. *Proceedings of the VLDB Endowment*, 5(3):241–252, 2011.
5. P. Heim, S. Lohmann, and T. Stegemann. Interactive relationship discovery via the semantic web. In *The Semantic Web: Research and Applications*, pages 303–317. 2010.
6. M. Nagarajan and et al. Predicting future scientific discoveries based on a networked analysis of the past literature. In *KDD*. ACM, 2015.
7. G. Pirrò. Explaining and suggesting relatedness in knowledge graphs. In *The Semantic Web-ISWC 2015*, pages 622–639. Springer, 2015.