

Improving Data Exploration in Graphs with Fuzzy Logic and Large-Scale Visualisation

Miguel Molina-Solana, David Birch, Yi-ke Guo

Data Science Institute, Imperial College London, UK

Abstract

This work presents three case-studies of how fuzzy logic can be combined with large-scale immersive visualisation to enhance the process of graph sensemaking, enabling interactive fuzzy filtering of large global views of graphs. The aim is to provide users a mechanism to quickly identify *interesting nodes* for further analysis. Fuzzy logic allows a flexible framework to ask human-like curiosity-driven questions over the data, and visualisation allows its communication and understanding. Together, these two technologies successfully empower novices and experts to a faster and deeper understanding of the underlying patterns in big datasets compared to traditional means in a desktop screen with crisp queries. Among other examples, we provide evidence of how these two technologies successfully enable the identification of relevant transaction patterns in the Bitcoin network.

Keywords: Graph sensemaking, Fuzzy Logic, Data Exploration, Large-scale Visualisation, Graph Visualisation

1. Introduction

With the advances of computational and sensing systems, huge amounts of data are created and stored every day. In order to gain useful knowledge from them, these data needs to be processed and interpreted in proper and meaningful ways. These tasks are still primarily performed by humans in possession of domain knowledge and the mathematical and statistical tools at their disposal.

In this context, Data Visualisation (together with descriptive analytics) has been proved an effective way of gaining insights into data [1], and is consistently the first and last step in most data science projects [2]. According to Leigh and colleagues [3], visualisation serves three important roles during the scientific discovery process: quick verification of correctness of simulation models during development, quick delivery of simulation results by close integration with the model, and easiness of understanding by lay audiences.

Fuzzy logic and fuzzy set theory [4] rely on the fact that human beings employ mostly words in their computing and reasoning processes. These are appropriate formalisms to handle imprecise knowledge, and particularly, imprecise queries over the data [5]. That kind of queries is even more relevant when performing exploratory data analysis over big datasets, and has its maximum exponent in the query “which are the most interesting pieces of data?”.

As it is well known, within the Fuzzy formalism, axioms and facts are not in general either true or false, but they

may hold to some degree of truth. Fuzzy Logic offers then a sound theoretical framework for curiosity-driven questions in the initial stages of the Data Exploration process.

In summary, both Data Visualisation and Fuzzy Logic improve interpretability and understanding of the data at hand, facilitating its query and exploration. In Chen’s words [6], “a visualisation process is a search process”. A search of meaningful insights than can drive further and deeper research and analysis on the dataset. Our thesis in this work is that blending these two technologies is a sensible proposal to intuitively perform an informed data exploration on big datasets, that would otherwise be very difficult, if not impossible, to process and fully understand. This combination has also the advantage to enable data understanding by broader lay audiences.

Describing datasets as graphs and networks, and relying on existing techniques to analyse them, has proven a successful line of action for plenty of scenarios. However, making sense of large graphs remains a fundamental challenge (reaching performance and usability issues), with few tools that allow users to interactively explore, visualise, and understand large graphs [7].

This paper precisely describes successful examples of such exploration employing fuzzy queries (particularly, looking for *interesting nodes*) and leveraging the *KPMG Data Observatory* at Imperial College’s Data Science Institute, the largest visualisation studio of its kind in Europe with 132 megapixels of display surface. The employment of a large-scale, high-resolution visualisation environment provides more effective data visualisation (with a global view) and greater insight into the data. By identifying interesting nodes, users are able to focus their attention

Email addresses: m.molina-solana@imperial.ac.uk
(Miguel Molina-Solana), david.birch@imperial.ac.uk
(David Birch), y.guo@imperial.ac.uk (Yi-ke Guo)

to those, which are in a more manageable scale than the original number of nodes in the large graph.

Among the three different case-studies presented, one focused on the Bitcoin network is currently allowing us to unveil unexpected transaction patterns and the observation of several evolving attacks [8]. Immersive large-scale visualisations and fuzzy queries allow rapid identification of *interesting* nodes, which sparked discussion and further investigation amongst researchers.

Accordingly, the main contributions of the paper are the following:

- We introduce a simple but effective framework for performing fuzzy queries (e.g. interesting nodes) over the data visualised in our large-scale visualisation environment at the Data Science Institute at Imperial College London.
- We present illustrative examples of high resolution visualisation of large graphs and how they can be explored using fuzzy queries to intuitively achieve a greater degree of understanding by the lay user.
- We discuss the advantages that the combination of those two technologies brings for Data Exploration, and graph sensemaking in particular.

The remainder of this paper is organized as follows. Section 2 provides some background on large-scale visualisation, graph sensemaking, and the concept of Fuzzy Data Exploration. Section 3 describes and discusses three examples of how our research group is using large data visualisation and fuzzy logic to gain better insights into large graphs. Section 4 discusses the advantages of our proposal for data exploration, and graphs in particular. The paper concludes by providing future lines of action and research.

2. Background

This section presents previous work on Large-scale Visualisation, Graph Sensemaking, and Fuzzy Data Exploration, three technologies that together can accelerate the process of gaining relevant insights into big datasets.

2.1. Data Visualisation

Visualisation is about communicating and perceiving data, both abstract and scientific, by means of the human visual system. Although it was not until the 1980s [9, 10], that data visualisation found a place in research, it has since then proved an effective way of gaining insights into data [1] both at the initial and final stages of the analysis.

Visual data exploration, in particular, is especially useful when little is known about the data and the exploration goals are vague. In this context, visual data exploration can be viewed as a hypothesis-generation process [1]. In it, the user can then interactively shift and adjust the analysis goals. In this process, hypotheses might get validated

or rejected on a visual basis, and new ones can be introduced. According to Shneiderman [11], visual data exploration follows a three-step process comprising overview, zoom and filter, and details-on-demand.

As datasets grow in complexity and scale, so should their visualisations, with many research teams and engineers developing hardware and software tools to enable large-scale visualisations on big screens and video-walls. The hypothesis is that more visual space will foster greater speed, accuracy, comprehension and confidence on the data analysis and interpretation. Pioneers on this topic were the works by the Electronic Visualization Laboratory at the University of Illinois at Chicago, and their CAVE [12] and CAVE2 [13] environments. Other examples of large visualisation environments reported in literature include [14, 15, 16, 3].

One of the advantages of high resolution visualisations is the ability to combine in the same display the aforementioned three steps (overview, zoom and filter and details-on-demand) described by Shneiderman. Besides the obvious technical difficulties that large visualisation poses (e.g. screen synchronization, high-resolution displays, bandwidth), it also prompts a shift from a single-user work flow towards a more collaborative one. This has critical advantages for the comprehension of data and identification of insights, and it is one of the key benefits of this type of systems.

Roberts et al. [17] recently reflected on the new challenges and opportunities that lie ahead for visualisation. Some of these challenges are precisely the ones addressed in our present work, mainly, large-scale immersive visualisation and human-like querying. They both aim at achieving a better comprehension of the growing datasets currently available, in intuitive ways for the users.

To perform this study, we leverage the large-scale visualisation observatory built at Imperial College London premises, which features a distributed rendering cluster with 64 46" HD screens. These screens have been arranged in a 6m diameter circle with 16 columns of 4 monitors. The 313 degree immersive space has a total resolution of 132M pixels across 37.31m².

2.2. Exploratory Data Analysis in Graphs

For graphs, like for any dataset, the first insights into them can be gained either through descriptive analytics or through visualisation. In that regard, Gephi [18] is probably the most well known piece of software for graph processing and exploration. Apart from real-time visualisation, this software is able to apply different layouts, apply filters, and calculate various metrics.

Data exploration in graphs (or *graph sensemaking* [7]) is not a new problem. The initial survey on graph visualisation techniques was authored by Herman et al. [19]; with Landesberger et al. [20] reviewing the works after 2000. More recently, Pienta et al. [7] organised most of those works according to their opportunities for understanding



Figure 1: Partial view of the *KPMG Data Observatory* at the Data Science Institute, Imperial College London. Its 64 46' screens create a 313 degree immersive space with a total resolution of 132M pixels. Users can walk in, and collaboratively explore data presented in different ways (charts, maps, animations, graphs...). The image shows Bitcoin transaction being displayed [8].

the underlying graph, providing a new angle into this area of research.

However, the increase in the relative size of datasets to account for real-world problems, has forced researchers and engineers to move to distributed and parallel proposals [21] in order to explore and process large graphs, and to calculate different measures over them. Machine learning research has made great advancements in developing scalable and distributed algorithms, but they often lack support for interactivity. Similarly, human-computer principles for interaction and visualisation are difficult to scale to the degree needed with the large size of current datasets.

For the sake of clarity, our proposal is not on improving the rendering of large graphs, or on improving the calculation of particular measures over them. For that, we rely on existing methods. Our proposal is to improve sensemaking by combining large scale visualisation and fuzzy queries on graph measures, with the aim of quickly identifying *interesting nodes* on global views of large graphs (which otherwise would be complicated and time-consuming), and informing latter stages of further research.

2.3. Fuzzy Data Exploration

Fuzzy Sets were initially proposed by Zadeh [4] to represent and manage imprecise and vague knowledge. Whilst in classical set theory elements either belong to a set or not, in fuzzy set theory elements can belong to a set to a certain degree.

Fuzzy set theory extends all classical set operations to fuzzy sets. The intersection, union, complement and implication set operations are performed by corresponding functions; respectively, a t-norm, a t-conorm, a negation, and an implication. The combination of them is called a fuzzy logic, and there are several of them depending on the selected functions [22].

For the sake of simplicity, in the rest of this article we will consider the fuzzy connectives originally proposed by Zadeh, namely the Gödel conjunction and disjunction, Lukasiewicz negation, and Kleene-Dienes implication (Table 1). They provide adequate results for our actual scenarios. Other typical fuzzy logics are Lukasiewicz, Gödel, and Product, which have different properties and interpretations [23].

Table 1: Specification of the Zadeh fuzzy logic

	Notation	Function
t-norm	$\alpha \otimes \beta$	$\min\{\alpha, \beta\}$
t-conorm	$\alpha \oplus \beta$	$\max\{\alpha, \beta\}$
negation	$\ominus \alpha$	$1 - \alpha$
implication	$\alpha \Rightarrow \beta$	$\max\{1 - \alpha, \beta\}$

Fuzzy logic has powered the methodology of *Computing with words* [24], a methodology in which words are used in place of numbers for computing and reasoning. It is a fact that human beings employ mostly words in their computing and reasoning processes, without needing an explicit use of any measurement. Fuzzy logic precisely enables this type of intuitive reasoning, and thus closes the gap between human reasoning and mathematical reasoning [25].

Fuzzy querying is not a new concept (see [26, 27] for early works, and [5, 28] for more recent ones), and it has been mainly applied as an extension of traditional querying in RDBS, enabling users to query them in natural language (inherently imprecise and vague). Our work can be considered as related with fuzzy querying because we will consistently ask for the *interesting nodes* within a graph. Fuzzy Logic will provide a powerful means of modelling the degree of interest of a node.

There have also been several previous proposals on visualising fuzzy rules and data [29, 30, 31, 32]. However, to the best of our knowledge, ours is the first attempt at applying fuzzy mechanisms in large-scale visualisation with the aim of highlighting relevant aspects, and improving the data exploration process. This fact is surprising taking into account that querying and filtering is one of the main tasks within visual data exploration.

To do the querying, we employ a *fuzzy rule-based system (FRBS)* in which fuzzy rules describe the quantitative relationship between variables in linguistic terms. As its crisp counterpart, FRBS comprises a *knowledge base* and an *inference engine*.

The *Knowledge base* collects the expert knowledge about the association of elements in the system. It is represented as a set of *IF-THEN* rules and *membership functions*. The current membership functions for the different linguistic tags and their rules have been defined by means of experts in a per-case basis (although they can be changed interactively if needed). In the future, we aim to automatically extract those functions from the current data being visualised, and to provide a means to aggregate different experts' definitions through consensus [33].

The *Inference Engine* is responsible of accepting the inputs after the fuzzification process, match them against the rules in the knowledge base, and provide the output values to the defuzzification module in order to obtain the output measurement. In our current implementation, we use Mamdani inference [34] due to its intuitiveness, and suitability for being understood by all users.

It should be noted that linguistic fuzzy inference offers advantages in terms of interpretability of the rules, although might present problems regarding its accuracy. This balance has been the subject of numerous discussions [35]. In our particular case (data exploration in big datasets), the loss of accuracy is perfectly acceptable measured against the gain in flexibility and intuitiveness that we achieve, specially in the early stages of the data exploration process.

3. Experimentation

To illustrate the power of large scale visualisation combined with fuzzy logic in order to perform effective data exploration of big datasets, we present in this section three examples focused on large graphs. In particular, we aim to highlight *interesting* nodes. In its abstract form, a graph is a representation of a set of objects and their interconnections. Graph models can be used to better understand the relations between different phenomena, such as traffic patterns, consumer behaviour, and weather anomalies, in addition to the analysis of social networks.

It is a fact, though, that graph analytics have been slowed down by complexity and scale. The huge number of interrelations in real-world graphs has prevented a deep analysis of these structures in appropriate manners. One of our lines of research within the Data Science Institute at Imperial College London is precisely on ways to better understand and communicate the information hidden within real-world graphs. Visualisation offers an intuitive mechanism to drive such an analysis, but yet, it is not enough as the human visual systems is unable to process information over certain thresholds.

There are several variables that can be defined over a graph and its elements. For illustrative purposes, we will only focus in this work on two of them (degree and weight of a node), but many others can be used (weight of an edge, graph diameter, average degree, ...). The two selected variables will be defined as linguistic variables (a variable whose values are not numbers but words or sentences in a natural or artificial language [25]), and their domains described over a set of three linguistic terms: *Low*, *Medium* and *High*. Linguistic terms are less specific than numerical values, but much more closer to the way humans express and use their knowledge.

The membership functions for each linguistic term will depend on each scenario, and so far, they are defined manually by a domain expert. In the rest of this section, along with the description of each example, the particular values

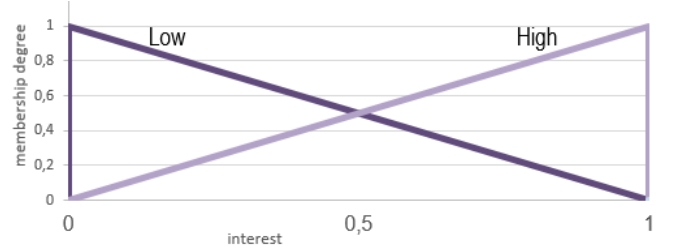


Figure 2: *Interest* linguistic variable with its corresponding linguistic terms.

for the corresponding membership functions will be provided. An appropriate definition of the linguistic terms is relevant as results will depend on them.

Another linguistic variable, called *interest* will be defined to characterise nodes in a graph, by means of two linguistic terms (*Low* and *High*). *Interest* is defined with the membership functions depicted in Figure 2.

In summary, the linguistic variables (with their corresponding linguistic term set) that we will consider for the examples in this section are the following:

Degree of Node = {Low, Medium, High}

Weight of Node = {Low, Medium, High}

Interest = {Low, High}

Applying fuzzy logic to graphs has been previously proposed in literature, mainly by proposing fuzzy graphs, and describing its modelling, visualisation and analytics [36]. Our current work, however, is focused on performing fuzzy queries on crisp graphs. As previously indicated, we will employ a Fuzzy Rule-based System to perform such inference, in order to identify what are the *interesting nodes*.

In terms of time complexity, and providing that we are not proposing any layout algorithm, our proposal for rendering a graph is linear in time as it requires visiting all nodes and links. Assuming that the crisp values for a metric (e.g. degree or weight) are already calculated for the nodes/links, the time to perform fuzzy queries over them varies linearly with the number of nodes (resp. links) in the large-scale graph (also with the number of rules in the knowledge base, but in practice, due to the different orders of magnitude, the number of rules can be dismissed).

Therefore, by performing the fuzzy calculation and rendering in the same operation (as it is the case), the time complexity is linear, and the difference between rendering a crisp query and a fuzzy one is imperceptible.

3.1. Network of relations

We will start with a simple case-study in order to demonstrate the potential of combining large-scale visualisation with fuzzy logic for data exploration purposes. In particular, we will consider a couple of networks of relations between entities.

Firstly, let's consider co-appearance network of characters in the novel *Les Misérables* [37], in which every character is a node and the coappearance in the novel is represented as a link. According to the previously defined linguistic variables, we will consider as interesting characters those with a high number of connections (i.e. nodes whose degree is high). In rule terms, *IF Degree IS High THEN Interest IS High*. We will also define as nodes with low interest those with a low degree of connections (i.e. *IF Degree IS Low THEN Interest IS Low*).

This graph, with the highlights, is shown in Figure 3. As expected, the green nodes (nodes with high interest) correspond to those with most connections in the graph and to the main characters in Victor Hugo's novel. Similarly, nodes with low interest (in red) correspond to nodes with a low degree. In this very simple example, the validity of the highlighting system can be easily checked in a visual way by plotting the links, even if no layout algorithm is applied.

The same query—interesting characters—can be performed (with the only modification of the particular values of the membership functions to adapt them to the new domain) to any other dataset. To demonstrate this, we made use of a larger one, the co-appearance network of characters in the Marvel heroes universe [38], in which every character is a node and the co-appearance in a comic is represented as a link. This graph contains 10,469 nodes and 224,126 edges.

The visual results of such a query are shown in Figure 4 with the interesting characters highlighted in yellow together with their names. All of them correspond to well-know heroes for the general public (among others, Spiderman, Hulk, Captain America, Thor, and Iron Man). Large-scale visualisation and fuzzy querying has allowed us (and any non-expert user) to quickly and visually check, over more than 10.400 nodes, that the most connected ones are indeed those corresponding to the famous heroes.

Figure 5 shows the membership functions for the different linguistic terms in variable 'degree' for the "Les Misérables" graph. Figure 6 shows the membership functions for that same linguistic variable in the "Marvel universe" scenario.

In both cases, and specially in the Marvel one, without a mechanism to identify *interesting nodes* that worth further investigation, it would be very difficult for a user to focus their research efforts on relevant aspects of the dataset. Our proposal provides a quick and natural way of querying for the most interesting nodes in these datasets. Note also that, despite no layout algorithm being applied, we are yet able to visually identify these nodes.

3.2. Discovering biggest contributors in Wikipedia

The second scenario to demonstrate the potential of intuitive fuzzy queries and large-scale visualization for gaining quick insights into the data is a graph of editions to the Wikipedia. In particular, the "Wikipedia Edits During the

Middle East Riots" dataset [39] from Elijah Meeks, while working as the digital humanities specialist at Stanford University.

This dataset has three types of nodes (topics, editors and editions) that we will render in different colours (see Figure 7). While the number of topics is contained, the number of editors and editions is high, making it infeasible to explore them all. Using our framework, we can again query for *interesting nodes*, which in this case are defined in two ways:

- *Interesting contributors* are those with a higher number of editions (i.e. more active). The linguistic variable is defined over the degree of an *editor* node.
- *Interesting editions* are substantial editions (i.e. large number of characters changed). They are defined over the weight of *edition* nodes.

Therefore, two rules exist: *IF Degree IS High THEN Interest IS High*, and *IF Weight IS High THEN Interest IS High*.

Figure 7 visually presents the dataset in our *Data Observatory*, with the three types of nodes rendered in different colours (topics in blue, editors in green, and editions in pink). *Interesting nodes* are also rendered (*interesting editors* in yellow, and *interesting editions* in red). As can be seen, the biggest editions are mostly related with the topic "Egyptian Revolution of 2011". Biggest contributors, on the other hand, are equally distributed across different topics, and no immediate trend is apparent.

For the sake of completeness of the descriptions, Figures 8 and 9 show the different trapezoidal membership functions for the linguistic variables involved in these example.

3.3. Discovering anomalous transactions in the Bitcoin network

Finally, we will explore the potential of fuzzy queries and large-scale visualization for Data Exploration in the dataset of Bitcoin's transactions, which is one of our current lines of research [8].

Bitcoin is the first implementation of a successful cryptocurrency system [40]. It takes the form of a peer-to-peer network in which the nodes interchanges bitcoins, saving all transactions (composed of 1..*n* outputs and 0..*n* inputs) into a distributed ledger known as the *blockchain*. By design, this database and its updates are public in order to allow a real-time majority consensus to form as to the current valid system state.

In a previous work [8], it was proposed a visualisation of transactions using a force-directed graph layout with gravity provided by the *SigmaJS* library and its *ForceAtlas2* continuous layout algorithm[41]. In it, a transaction is visualised as a single vertex, associated by edges to its component orange input and blue output vertexes. Independent transactions are visually associated to each other in two ways: either directly through an existing output

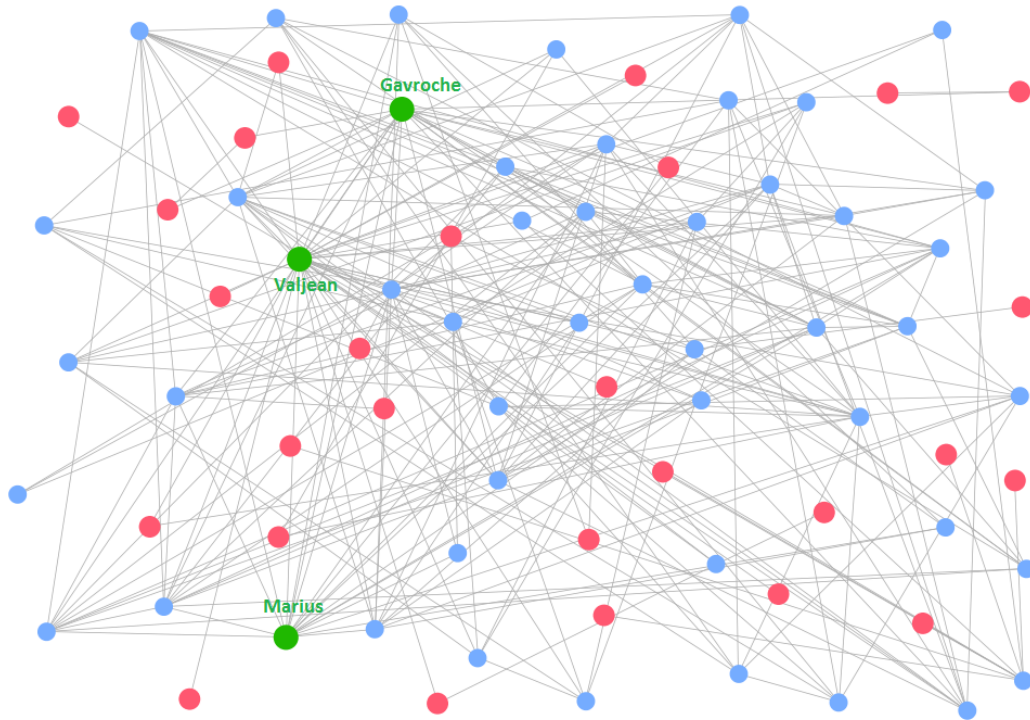


Figure 3: Highlighted in green are the interesting characters (defined as those with *high* degree of connections) in the *Les Misérables* graph. Highlighted in red are the characters with a *low* number of connections. Even without a proper layout, it is easy for users to visually identify which are a few nodes that might have some further interest.

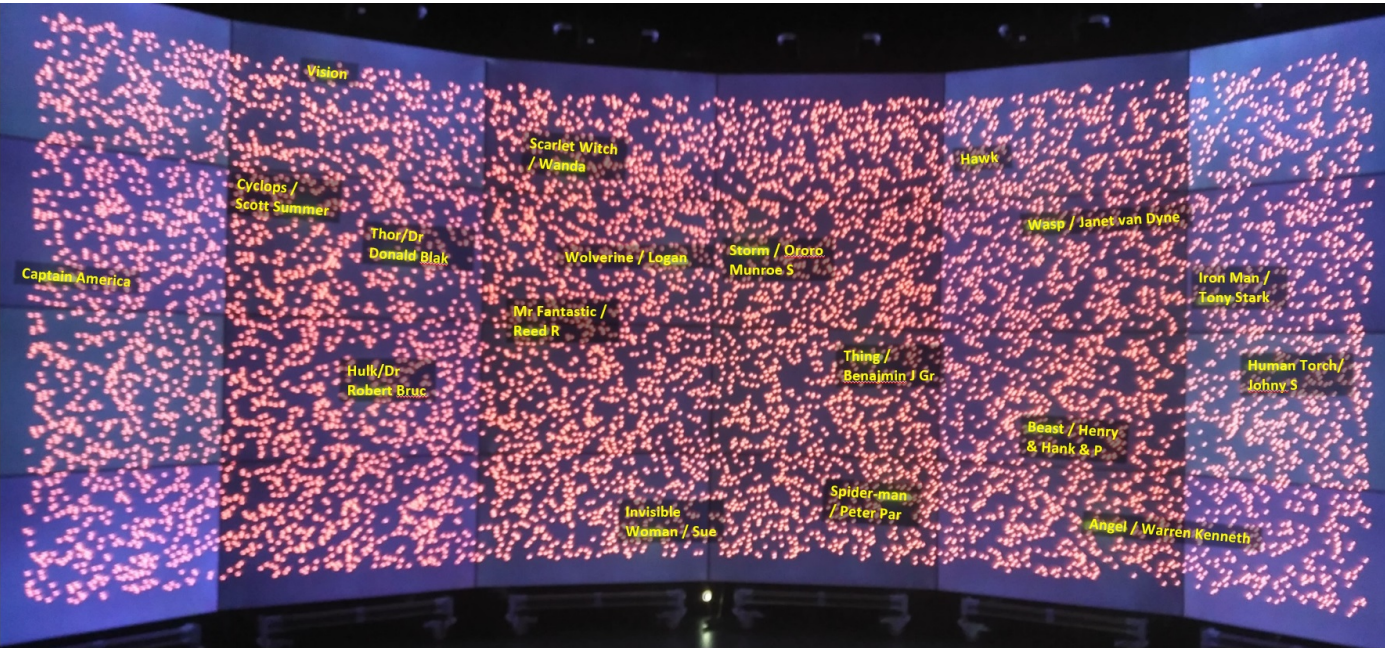


Figure 4: Highlighted in yellow are the *interesting* characters (defined as those with *high* degree of connections) in the Marvel universe. A random layout is applied, and links have been removed for clarity. To improve legibility, labels have been added manually on top of the image.

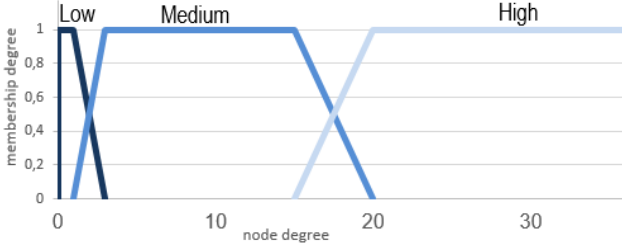


Figure 5: For *Les Miserables* example, definition of the *Degree* linguistic variable with its associated linguistic terms.

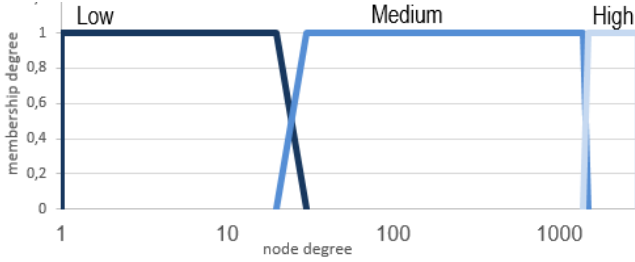


Figure 6: For *Marvel Universe* example, definition of the *Degree* linguistic variable with its associated linguistic terms. Note the logarithmic scale for the domain.

becoming an input to a new transaction within the time-frame of the visualisation, or indirectly through the re-use of the same cryptographic public key within an element of a transaction, which we connect with a grey edge.

In the context of this visualisation, a normal human-like Bitcoin transaction is composed of one input and two outputs (the amount paid and the change back). However, looking at the visualisation, it is immediately evident that there are some transactions that do not look human-like, as they have plenty of links and interconnections. For more details on some of these patterns, we encourage the reader to check [8].

Fuzzy logic allows us to describe human-like transactions as those with a *small* number of inputs and a *small* number of outputs. Note that, the term *small* is used in this query and sound perfectly normal for a human being. However, in terms of computer interaction, *small* should be translated into an exact value, and that is not generally an easy task for non-experts when first approaching the data.

The power of fuzzy querying is fully exploited when more complex queries are performed such as: “transactions with a *high* number of outputs, and a *low* value” (which we hypothesise correspond to malicious transactions trying to spam the network with high-frequency, complex, low-value transactions). In this query, two linguistic variables are involved. In rule terms, it would be expressed as *IF Degree IS High AND Weight IS Low THEN Interest IS High*.

This query has been applied to a graph containing all the transactions within three consecutive Bitcoin blocks (particularly, blocks #364132, #364133 and #364134). This graph comprises more than 37.000 nodes and almost

50.000 edges.

A visualisation of *interesting* transactions (that we consider malicious) over that graph is presented in Figure 10. It confirms that the worm-like structures are certainly suspicious in its nature. Of course, these insights should be later further explored and confirmed by deeper analytics, but in a very intuitive way, visualisation and fuzzy querying have put us on our way to do so.

For the sake of completeness of the descriptions, Figures 11 and 12 show the different trapezoidal membership functions for the linguistic variables involved in the Bitcoin example.

4. Discussion

The previous examples and use cases show that the proposed combination of large scale visualisation with fuzzy queries notably increases the capabilities and power of Data Exploration, providing support for much more expressive insights into the data. Fuzzy queries retrieve not only values explicitly asserted in the data, but also information automatically inferred by a logic-based reasoning process. Therefore, Fuzzy Logic offers a sound theoretical framework for curiosity-driven questions in the initial stages of the Data Exploration process. The availability of languages and tools facilitates the implementation of this kind of solutions.

In its current implementation, linguistic terms have to be defined on a per-case basis (although they can be easily modified, if required). The results (*interesting nodes*) are directly dependant on the definition of the linguistic tags and the rules. Interestingly enough, for our purposes of Data Exploration they do not need to be very accurate, and a rough idea of their values is generally sufficient to achieve meaningful results.

For the cases demonstrated in this paper (with different combinations of variables and rules), this framework is sufficient for identifying the *interesting nodes* in large graphs, greatly reducing the number of nodes a user has to pay attention to. In this context, reducing the core of the linguistic tags that lead to high degrees of interest would mean that a smaller number of nodes qualifies to be consider as *interesting*.

The size and detail that the Observatory environment enables in visualisations is the key to gain insights into the data in a much quick and effective way. At least, that is the general comment of the users at our environment, and the fact reported in several other works [16, 3]. An environment such as ours enables Shneiderman’s three-step process for visual data exploration (overview, zoom and filter, and details-on-demand) to happen in an interactive iterative and social means.

There are as well extensive reports in the literature on the benefits of the *Computing with words* paradigm for Human Computer interaction and querying. Our work has demonstrated that it can nicely blend together with

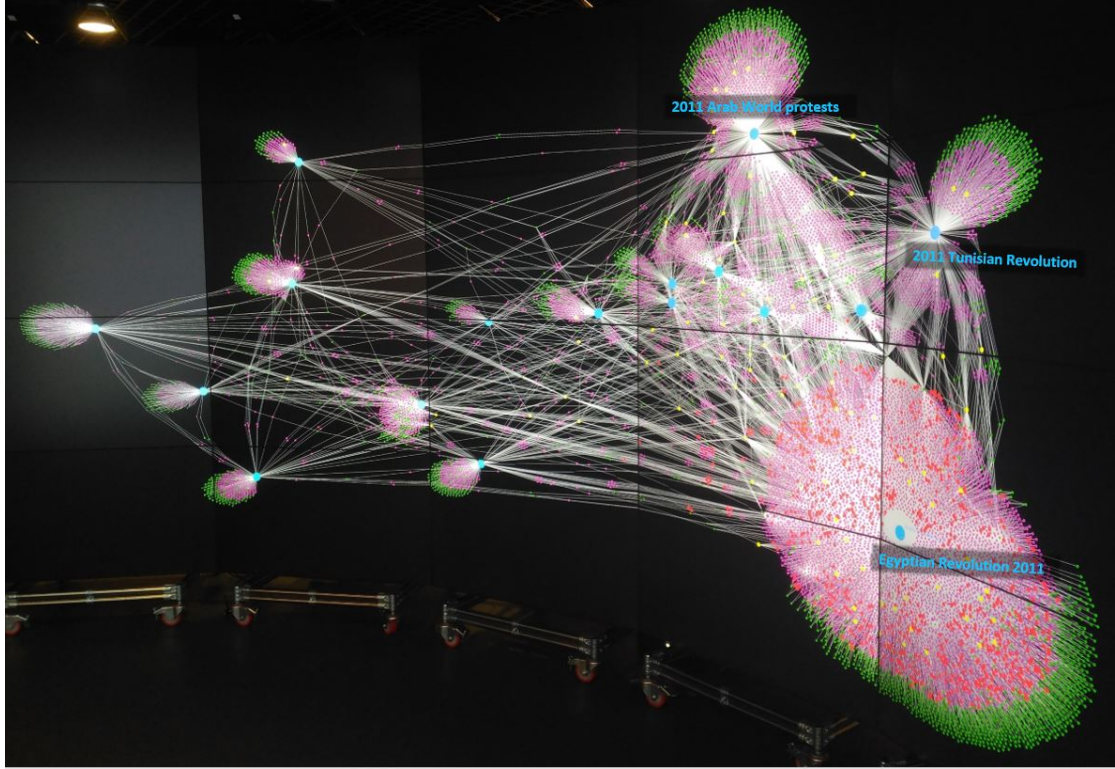


Figure 7: Large scale visualisation at the *KPMG Data Observatory* of a graph representing Wikipedia edits during the Middle East riots. Topics are represented as blue nodes, editors are green, and editions are pink. *Interesting contributors* (more active) are highlighted in yellow, whilst *interesting editions* (more substantial) are highlighted in red. Label of three topics have been manually included

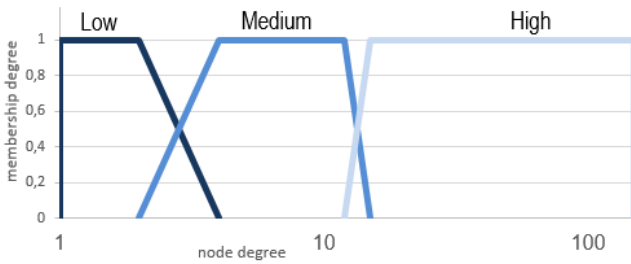


Figure 8: For Wikipedia Riots example, definition of the *Degree* linguistic variable and its associated linguistic terms

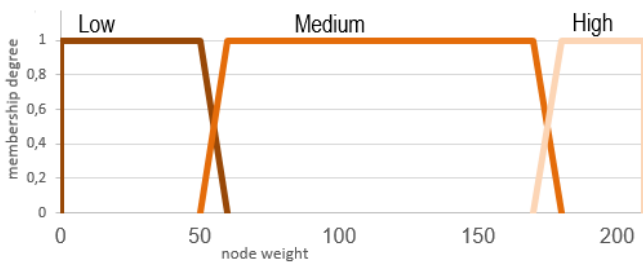


Figure 9: For Wikipedia Riots example, definition of the *Weight* linguistic variable with its associated linguistic terms

large scale visualisation, allowing a smoother and more intuitive data exploration process over big datasets for lay audiences and experts alike.

One of the proclaimed features of big data analysis is that, as the amount of data is so big, patterns are resilient to random errors and they do not need to be very accurate; just enough to point trends out.

In the same way, the accuracy of the results does not need to be high for the initial stages of Data Exploration in large-scale visualisations. In fact, what is needed at this point is a high level and general understanding of the whole data, and the ability to quickly spot relevant patterns (in our own words, *interesting nodes*). Both fuzzy querying and large-scale visualisation precisely contribute to that.

The additional advantage of a facility such as the *KPMG Data Observatory* (which we have not specially remarked in this paper) is its social facet which naturally fosters the interactions between teams, and facilitates the collaborative evaluation and exploration of graphs and patterns. Joint data exploration by members of a team is another way of gaining deeper insights on the data, compared with doing it alone.

Finally, and although we have shown here the application to graph exploration, other visualisations can perfectly benefit from fuzzy querying, and we are indeed currently exploring these possibilities.

For instance, geolocated data can be queried in terms

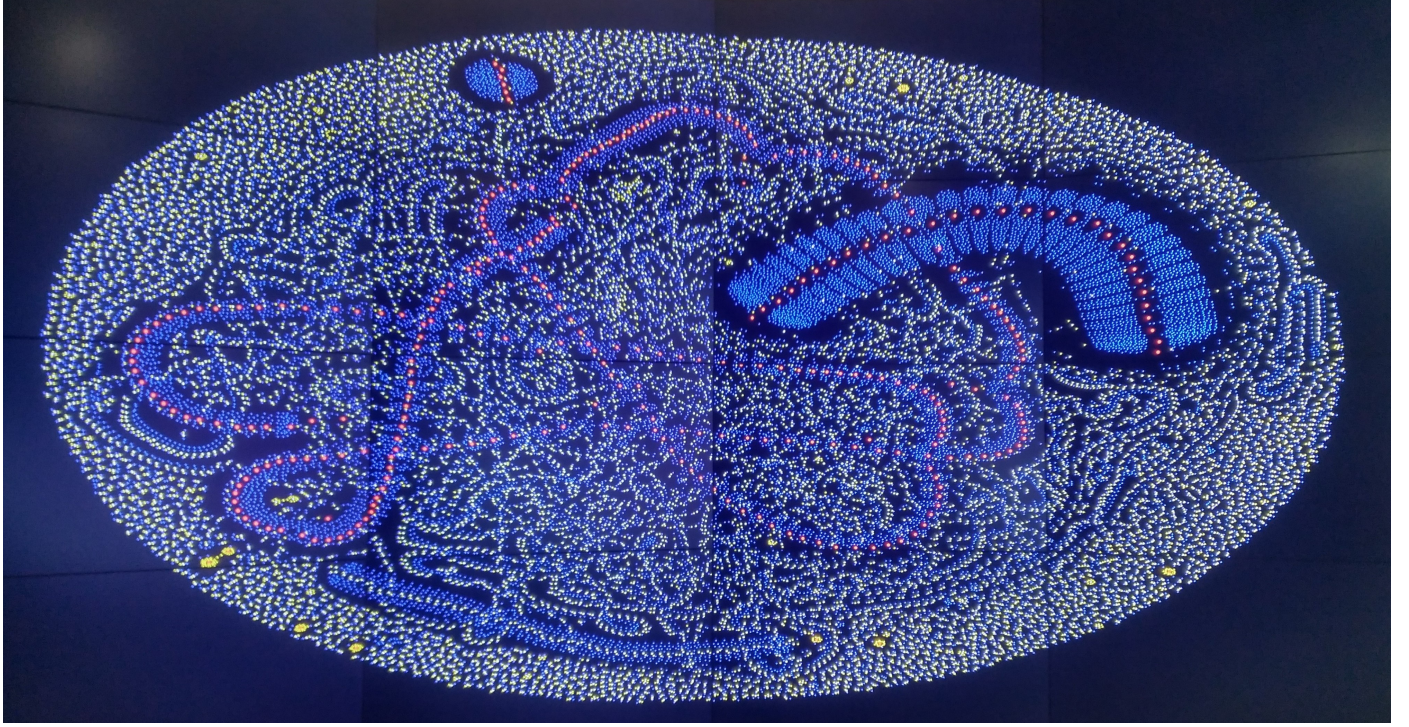


Figure 10: Large scale visualisation at the *KPMG Data Observatory* of a graph representing Bitcoin transactions. *Interesting transactions* (which we consider as malicious) are highlighted in red. Once visually identified, they can be further investigated and processed, as we did in [8].

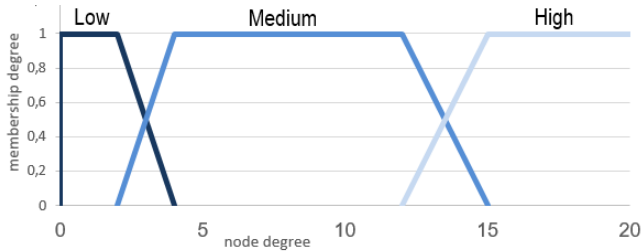


Figure 11: For Bitcoin example, definition of the *Degree* linguistic variable and its associated linguistic terms.

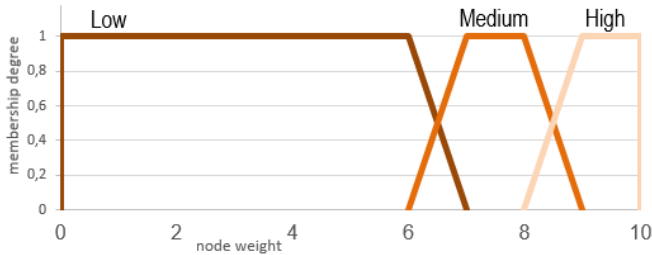


Figure 12: For Bitcoin example, definition of the *Weight* linguistic variable and its associated linguistic terms.

of approximate distance to different reference points. A query such as “Most crowded underground stations close to a hospital” is certainly possible within the fuzzy formalism (with *most crowded* and *close to* as fuzzy terms), and nicely represent the kind of query a manager might ask. It is certainly more natural than “Stations with more than 1500 passengers per hour and located less than 100 metres to a hospital”.

The previous fuzzy query can be further extended with temporal information (which can again be expressed in fuzzy terms): “Most crowded underground stations in the mornings, close to a hospital”. Certainly, the possibilities are endless.

5. Conclusions and further work

This paper has illustrated how the combination of large-scale high resolution visualisation environments with fuzzy logic can provide a valuable tool to gain understanding of and insight into big data by means of very simple human-like queries. This has been shown through the exploration of high-resolution visualisations of three different graphs representing very different information.

These visualisations (some of them of over 30.000 vertices) together with fuzzy queries on them, provide a quick high-level overview and drilled-down detail of interesting knowledge within the data. In particular, with a very simple formalism, we were able to provide users a way

to quickly identify *interesting* nodes, which might be defined differently for different domains. Once those nodes are identified, further research can be done around them and their properties.

Although the fuzzy models employed here are far from complex, the reader can see the interesting results achieved, and the huge possibilities ahead if more complex queries are required for a given domain. Both fuzzy query and large-scale visualisation empower a more intuitive, human-friendly way of gaining insights into the data in big datasets.

Of course, the definition of an *interesting node* varies for each scenario, and is a difficult problem in itself. To define these, we relied on human experts to provide relevant linguistic terms for each one of the related variables. We are looking into ways of automatising this definition.

This work has concentrated on the visualisation of fuzzy queries (and particularly, on *interesting nodes*) over crisp graphs. A further step will be the visual representation of fuzzy graphs (i.e. graphs in which the weight of nodes and edges are themselves fuzzy), and fuzzy information. In those scenarios, fuzzy querying can be directly applied without requiring any fuzzification process.

In addition to that, we are currently extending the fuzzy querying functionality to other visualisations of data in the environment. The final goal is to integrate fuzzy querying as a natural mechanism to gain insights into the visualised data, in the same way that crisp querying, filtering and zooming are nowadays used.

We share with Roberts et al. [17] the strong believe that data visualisation is “the next big thing”. But we also believe that Fuzzy Logic has its role to play on it. As we have demonstrated, the potential certainly exists.

Acknowledgments

The infrastructure for the *Data Observatory* has been partially funded by KPMG and Imperial College London. The authors would like to acknowledge the work by Dan McGinn and Suryansh Rastogi in preparing the Bitcoin data, as well as Min Hua Chia who developed the software for visualising graphs in the environment.

- [1] D. A. Keim, Visual exploration of large data sets, *Communications of the ACM* 44 (8) (2001) 38–44. doi:10.1145/381641.381656.
- [2] U. Fayyad, G. G. Grinstein, A. Wierse (Eds.), *Information Visualization in Data Mining and Knowledge Discovery*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2002.
- [3] J. Leigh, A. Johnson, L. Renambot, T. Peterka, B. Jeong, D. J. Sandin, J. Talandis, R. Jagodic, S. Nam, H. Hur, Y. Sun, Scalable resolution display walls, *Proceedings IEEE* 101 (1) (2013) 115–129. doi:10.1109/JPROC.2012.2191609.
- [4] L. Zadeh, Fuzzy sets, *Information and Control* 8 (3) (1965) 338–353.
- [5] J. Kacprzyk, S. Zadrozny, G. D. Tré, Fuzziness in database management systems: half a century of developments and future prospects, *Fuzzy Sets and Systems* 281 (2015) 300–307. doi:10.1016/j.fss.2015.06.011.
- [6] M. Chen, D. Ebert, H. Hagen, R. Laramée, R. van Liere, K.-L. Ma, W. Ribarsky, G. Scheuermann, D. Silver, Data, information, and knowledge in visualization, *IEEE Computer Graphics and Applications* 29 (1) (2009) 12–19. doi:10.1109/MCG.2009.6.
- [7] R. Pienta, J. Abello, M. Kahng, D. H. Chau, Scalable graph exploration and visualization: Sensemaking challenges and opportunities, in: *2015 International Conference on Big Data and Smart Computing (BIGCOMP)*, 2015, pp. 271–278. doi:10.1109/35021BIGCOMP.2015.7072812.
- [8] D. McGinn, D. Birch, D. Akroyd, M. Molina-Solana, Y. Guo, W. J. Knottenbelt, Visualizing dynamic bitcoin transaction patterns, *Big Data* 4 (2) (2016) 109–119. doi:10.1089/big.2015.0056.
- [9] E. R. Tufte, *The Visual Display of Quantitative Information*, 2nd Edition, Graphics Press, 2001.
- [10] B. H. McCormick, Visualization in scientific computing, *ACM SIGBIO Newsletter* 10 (1) (1988) 15–21. doi:10.1145/43965.43966.
- [11] B. Shneiderman, The eyes have it: a task by data type taxonomy for information visualizations, in: *Procs. 1996 IEEE Symposium on Visual Languages*, 1996, pp. 336–343. doi:10.1109/VL.1996.545307.
- [12] C. Cruz-Neira, D. J. Sandin, T. A. De Fanti, R. V. Kenyon, J. C. Hart, The CAVE: Audio visual experience automatic virtual environment, *Communications of the ACM* 35 (6) (1992) 64–72. doi:10.1145/129888.129892.
- [13] A. Febretti, A. Nishimoto, T. Thigpen, J. Talandis, L. Long, J. Pirtle, T. Peterka, A. Verlo, M. Brown, D. Plepys, D. Sandin, L. Renambot, A. Johnson, J. Leigh, CAVE2: A hybrid reality environment for immersive simulation and information analysis, in: *Proc. IS&T / SPIE Electronic Imaging, The Engineering Reality of Virtual Reality 2013*, San Francisco, US, 2013, p. 864903. doi:10.1117/12.2005484.
- [14] G. Wallace, O. J. Anshus, P. Bi, H. Chen, Y. Chen, D. Clark, P. Cook, A. Finkelstein, T. Funkhouser, A. Gupta, M. Hibbs, K. Li, Z. Liu, R. Samanta, R. Sukthankar, O. Troyanskaya, Tools and applications for large-scale display walls, *IEEE Computer Graphics and Applications* 25 (4) (2005) 24–33. doi:10.1109/MCG.2005.89.
- [15] K. Li, M. Hibbs, G. Wallace, O. Troyanskaya, Dynamic scalable visualization for collaborative scientific applications, in: *Proc. 19th IEEE Int. Parallel and Distributed Processing Symposium (IPDPS’05)*, IEEE Computer Society, 2005, pp. 225.2–. doi:10.1109/IPDPS.2005.183.
- [16] K.-U. Doerr, F. Kuester, CGLX: a scalable, high-performance visualization framework for networked display environments, *IEEE Trans Vis Comput Graph* 17 (3) (2011) 320–332. doi:10.1109/TVCG.2010.59.
- [17] J. Roberts, P. Ritsos, S. Badam, D. Brodbeck, J. Kennedy, N. Elmquist, Visualization beyond the desktop—the next big thing, *IEEE Computer Graphics and Applications* 34 (6) (2014) 26–34. doi:10.1109/MCG.2014.82.
- [18] M. Bastian, S. Heymann, M. Jacomy, Gephi: An open source software for exploring and manipulating networks, in: *Proc. International AAAI Conference on Weblogs and Social Media*, 2009. doi:10.13140/2.1.1341.1520.
- [19] I. Herman, G. Melancon, M. Marshall, Graph visualization and navigation in information visualization: A survey, *IEEE Transactions on Visualization and Computer Graphics* 6 (1) (2000) 24–43. doi:10.1109/2945.841119.
- [20] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, D. Fellner, Visual analysis of large graphs: state-of-the-art and future research challenges, *Computer Graphics Forum* 30 (6) (2011) 1719–1749. doi:10.1111/j.1467-8659.2011.01898.x.
- [21] D. S. Banerjee, A. Kumar, M. Chaitanya, S. Sharma, K. Kothapalli, Work efficient parallel algorithms for large graph exploration on emerging heterogeneous architectures, *Journal of Parallel and Distributed Computing* 76 (2015) 81–93. doi:10.1016/j.jpdc.2014.11.006.

- [22] H. Bustince, E. Barrenechea, M. Pagola, J. Fernandez, Z. Xu, B. Bedregal, J. Montero, H. Hagraas, F. Herrera, B. De Baets, A historical account of types of fuzzy sets and their relationships, *IEEE Transactions on Fuzzy Systems* 24 (1) (2016) 179–194. doi:10.1109/TFUZZ.2015.2451692.
- [23] G. J. Klir, B. Yuan, *Fuzzy Sets and Fuzzy Logic: Theory and Applications*, Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1995.
- [24] L. A. Zadeh, Fuzzy logic = computing with words, *IEEE Transactions on Fuzzy Systems* 4 (2) (1996) 103–111. doi:10.1109/91.493904.
- [25] L. Zadeh, The concept of a linguistic variable and its application to approximate reasoningII, *Information Sciences* 8 (4) (1975) 301–357. doi:10.1016/0020-0255(75)90046-8.
- [26] S.-K. Chang, J.-S. Ke, Translation of fuzzy queries for relational database system, *IEEE Trans. Pattern Anal. Mach. Intell.* 1 (3) (1979) 281–294.
- [27] J. Kacprzyk, A. Ziolkowski, Database queries with fuzzy linguistic quantifiers, *IEEE Transactions on Systems, Man and Cybernetics* 16 (3) (1986) 474–479. doi:10.1109/TSMC.1986.4308982.
- [28] F. Albertetti, L. Grossrieder, O. Ribaux, K. Stoffel, Change points detection in crime-related time series: An on-line fuzzy approach based on a shape space representation, *Applied Soft Computing* 40 (2016) 441–454. doi:10.1016/j.asoc.2015.12.004.
- [29] R. Brown, B. Pham, Visualisation of fuzzy decision support information: a case study, in: *Proc. 12th IEEE Int. Conf. Fuzzy Systems (FUZZ '03)*, Vol. 1, 2003, pp. 601–606. doi:10.1109/FUZZ.2003.1209432.
- [30] B. Pham, R. Brown, Visualisation of fuzzy systems: requirements, techniques and framework, *Future Generation Computer Systems* 21 (7) (2005) 1199–1212. doi:10.1016/j.future.2004.04.007.
- [31] T. Doan, A. Hamilton-Wright, Visualization of assertion confidence in fuzzy rules, in: *Fuzzy Information Processing Society, 2009. NAFIPS 2009. Annual Meeting of the North American*, 2009, pp. 1–5. doi:10.1109/NAFIPS.2009.5156425.
- [32] C. Vehlow, T. Reinhardt, D. Weiskopf, Visualizing fuzzy overlapping communities in networks, *IEEE Trans. Vis. Comput. Graphics* 19 (12) (2013) 2486–2495. doi:10.1109/TVCG.2013.232.
- [33] G. Bordogna, M. Fedrizzi, G. Passi, A linguistic modelling of consensus in group decision making based on OWA operators, *IEEE Transactions on Systems, Man, and Cybernetics* 27 (1997) 126–132. doi:10.1109/3468.553232.
- [34] E. Mamdani, Application of fuzzy logic to approximate reasoning using linguistic synthesis, *IEEE Trans. Comput.* C-26 (12) (1977) 1182–1191. doi:10.1109/TC.1977.1674779.
- [35] J. Casillas, O. Cordon, F. Herrera, L. Magdalena (Eds.), *Interpretability issues in Fuzzy Modeling*, Springer-Verlag Berlin Heidelberg, 2003. doi:10.1007/978-3-540-37057-4.
- [36] O. Pivert, G. Smits, V. Thion, Expression and efficient processing of fuzzy queries in a graph database context, in: *Proc. 2015 IEEE Int. Conf. Fuzzy Systems (FUZZ-IEEE)*, 2015, pp. 1–8. doi:10.1109/FUZZ-IEEE.2015.7337849.
- [37] D. E. Knuth, *The Stanford GraphBase: A Platform for Combinatorial Computing*, Addison-Wesley, Reading, MA, 1993.
- [38] R. Alberich, J. Miro-Julia, F. Rosselló, Marvel universe looks almost like a real social network, eprint at [arXiv:cond-mat/0202174](https://arxiv.org/abs/cond-mat/0202174).
- [39] Wikipedia edits during the Middle East riots, <https://dhs.stanford.edu/dh/gephi/wikiriots.gexf>, accessed: 2016-01-26.
- [40] S. Nakamoto, *Bitcoin: A Peer-to-Peer Electronic Cash System*, Tech. rep., available at <https://bitcoin.org/bitcoin.pdf> Accessed on 2016-09-02 (2008). URL <https://bitcoin.org/bitcoin.pdf>
- [41] M. Jacomy, T. Venturini, S. Heymann, M. Bastian, ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software, *PLoS ONE* 9 (6) (2014) e98679. doi:10.1371/journal.pone.0098679.