DQNET: Assessment of Quality Regulation System as Complex Information Network

Tamás Csiszér

Óbuda University, Bécsi út 96/b, 1034 Budapest, Hungary
csiszer.tamas@rkk.uni-obuda.hu

Abstract: This article introduces a set of indicators and their interpretation called DQNET for the assessment of information structures in documents of quality regulatory systems. This complex system is considered a network with a piece of information in documents nodes; and links between them arcs. Like in citation network of scientific publications there are several network indicators in such information networks, which can reflect the 'positions' and 'roles' of elements in this system. By in- and out-degrees and other matrices documentations can be identified with, e.g. 'high importance' or with 'high sensitivity', requiring different ways of handling. By the indicators of structure functional suitability of regulation can be analyzed and predicted too.

Keywords: Documentation analyses; importance and sensitivity; networkscience

1 Introduction

Grouping or clustering documentation by calculating the similarity or the distance of documents or of their parts as the entities of regulation systems are one of the most important fields of the research of complex information networks.

Many scientists proposed indicators of document similarities, focusing on different elements of documents like words and phrases. One of these researchers, Wang proposes a method to represent a document as a typed Heterogeneous Information Network (HIN), where the entities and relations are annotated with types [5]. He and his colleagues underline that most of researches in the field of documentation networks are focusing on similarities between documents and do not put enough efforts on links sourced in heterophily, i.e. the difference between documents [7]. Yang proposes hierarchical attention network for classifying documents according to its hierarchy and the importance of content (word, sentence and document vectors) [6]. Tan presents the latent quality model (LQM). LQM associates each document with a latent quality score, which provides a measure of the impact or popularity of a document [3]. Wan proposes Cluster-based Conditional Markov Random Walk Model (ClusterCMRW) and the

Cluster-based HITS Model (ClusterHITS) to find parts of different documentations related to the same content to summarize information [4]. Cao developed a Ranking framework upon Recursive Neural Networks to rank sentences for multi-document summarization [1]. Carley applies Dynamic Network Analysis (DNA) approach to create and analyze multi-mode and multi-link networks [2].

All of these approaches were involved into the development process of DQNET. Documentation systems consist of many elements such as manuals, descriptions of procedures and products, forms, templates and others, published on paper or in electronic format. There is a huge number of links among their parts indicating the connections of regulations. One can find regulation holes and redundancy too. Due to this complexity, these systems are difficult to create, maintain, assess, upgrade and improve, so these activities should be supported by analytical and development methods based on qualitative and quantitative measurements. The purpose of these methods is to give evidence of proper or improper structure of documentation or – in general – information systems. In the following chapters, we introduce a set of indicators of DQNET that can represent the internal and external properties of the elements of such kind of systems.

2 Theory

2.1 Structure

A network-like representation of a documentation system can be seen in Figure 1. General nodes (black dots) represent the elements of the system. Links between documents are represented by grey arrows, indicating the direction of links too. There can be seen some special types or groups of nodes as well, represented by colored dots and circles as follows:

- Blue dots: reversed regulation. It may show the problem of regulating something with a link to another document, which has a link the other way around. It may be useful if these links belong to the same parts of documents and indicate the two directions of the same connection, or if these links belong to different parts of documents and indicate different connections, but it may indicate that these documents are linked to each other with a regulation hole. In these cases, the higher the number of links between two documents is, the stronger the connection of them can be detected.
- Purple dots: regulation loop. It may show the similar property of regulation described above, with involving more than two documents.

Such kinds of grouping of documents should be assessed according to the same approach.

- Orange dots: regulation chain. It shows how a rule defined with a set of documents with one-direction links. Obviously, it should be considered as chain if their links belong to the same parts of documentations.
- Red circles: hubs. These documents may have important roles in the system due to their links to other elements.
- Green circles: regulation trees. One document links to more than one. Changes of this document can have huge influence on others or vice-versa.
- Blue circles: regulation islands. These have noconnection to the other parts of the regulation network.

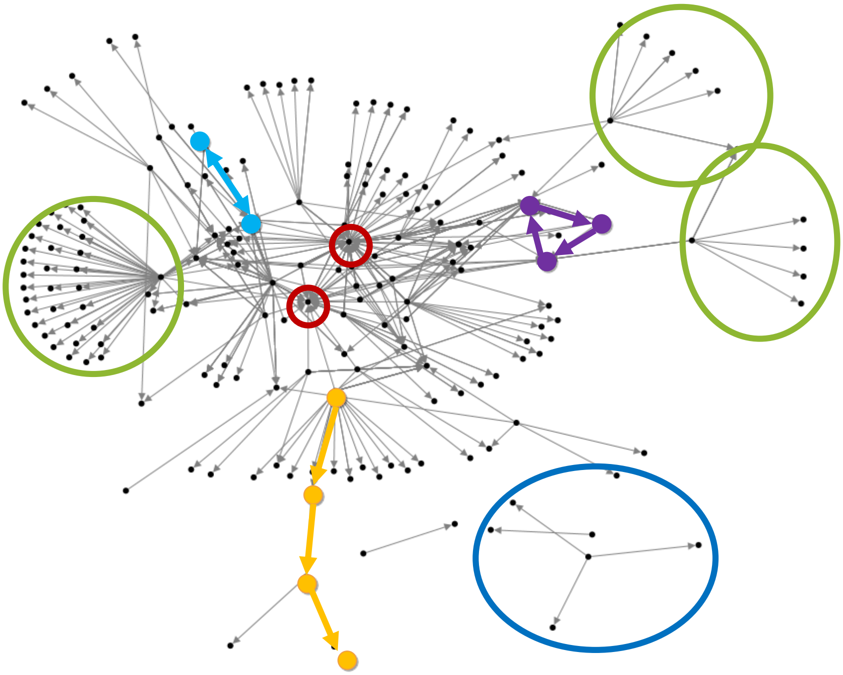

Figure 1

An example for the network-like implementation of documentation systems

Networks can be described with the number and structure of these special types and groups of nodes.

2.2 Internal Properties

Internal properties of documents determine how easy it is to understand, memorize and apply regulations described in documentation system. Most important related indicators:

- Size related indicators: number of pages, words, sentences and lines.
- Sentence structure related indicators: length of sentence and words, rate of number of words and sentences, rate of number of commas and sentences, rate of long sentences.
- Text structure related indicators: rate of number of sentences and paragraphs, rate of number of paragraphs and pages.
- Document structure related indicators: number of appendices and chapters.

With these indicators documents can be qualified from different perspectives as follows:

- Understandability: how easy itis to understand regulations.
- Notability: how easy it is to memorize regulations.
- Accountability: how easy it is to identify responsibilities.
- Searchability: how easy it is to find user or case relevant information.
- Applicability: how easy it is to apply the regulations during operation.

2.3 External Properties

External properties can be described by well-known network indicators as follows:

- In-degree - Importance: number of incoming links of a nod. The higher the in-degree of a nod is, the more important the document represented by the nod is.
- Out-degree – Sensitivity: number of outgoing links of a nod. The higher the out-degree of a nod is, the more sensitive the document represented by the nod is.
- Degree distribution – Evenness: how links are distributed to nods. It shows how evenly nods are connected to each other. Some questions that can be answered by this indicator: 1) Can a chain of links among all documents be found?; 2) Are there isolated elements or groups in network?; 3) Are there big differences among the degrees of nods?

- Betweenness – Criticality: in what part of the shortest paths between any pair of nods the associated nod takes part. The higher the betweenness is, the more critical role the document has in the network.
- Closeness – Simplicity: how close nods are to each other. The shortest the average distance (number of links on the path) among the nods is, the simpler the network is. It may help us to make the system of connections simpler.
- Clustering coefficient – Looping: rate of realized and possible numbers of triangles of nodes. It shows how many connected circles of 3 documents have been created.
- Reciprocated vertex pair ratio – Reciprocity: rate of two-directional to one-directional links of nods.

Knowing the values of network indicators, documentation network can be qualified. Some examples of qualification:

- Clarity: documents are connected to each other precisely; sender and receiver documents of the links can be identified exactly.
- Relevance: links connect the proper parts of proper documents.
- Redundancy: two-directional links between two documents are not redundant, i.e. indicate two different connections.
- Contradiction: rules defined in connected documents are consistent.
- Completeness: regulation hole cannot be found.

2.4 Network-based Optimizationof Documentation System

There are several ways to optimize a documentation system. It depends on the goals, organization structure and culture, skills of users, technical environment, level of automatization, etc. Due to this complexity there is no single ideal solution, but some important features can be defined based on the properties described above.

One of the fundamental goals of creating documents is to define regulation for operation, which must be easy to find, understand, memorize and apply. It can be ensured if rules for conducting a particular activity are handled as individual information package represented by only one nod in the regulation network. This information package consists of short sentences and graphical elements. Two or more nods are connected by links if activities represented by these nods 1) form a predecessor-successor pair of process steps, 2) are allocated to the same equipment, 3) need the same human skills to be done, etc. Nods and different types of links among them form different networks of operation rules. Different subgraphs of these networks belong to processes, products, resources, organization groups and localizations. According to the grouping principle, different types of documentation (e.g. process manuals, product descriptions, etc.) can be created too. An example of this rule-based network can be seen in Figure 2.

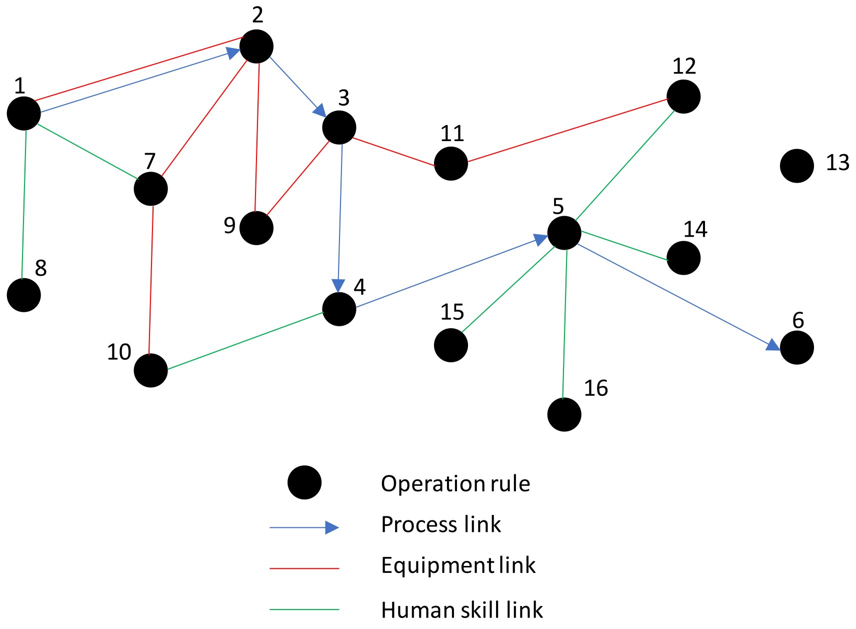

Figure 2
Example for rule-based network (part of whole network)

3 Case Study

An international financial organization has a complex system of documentations. Due to the order of Central Regulatory Office process documentations have to be modified to meet new requirements. The management decided to analyze the documentation structure with DQNET network indicators. The associated graphs and calculations are generated by NodeXL application.

3.1 Overall Metrics

The whole system can be seen in Figure 3. The Overall metrics are presented inTable 1.

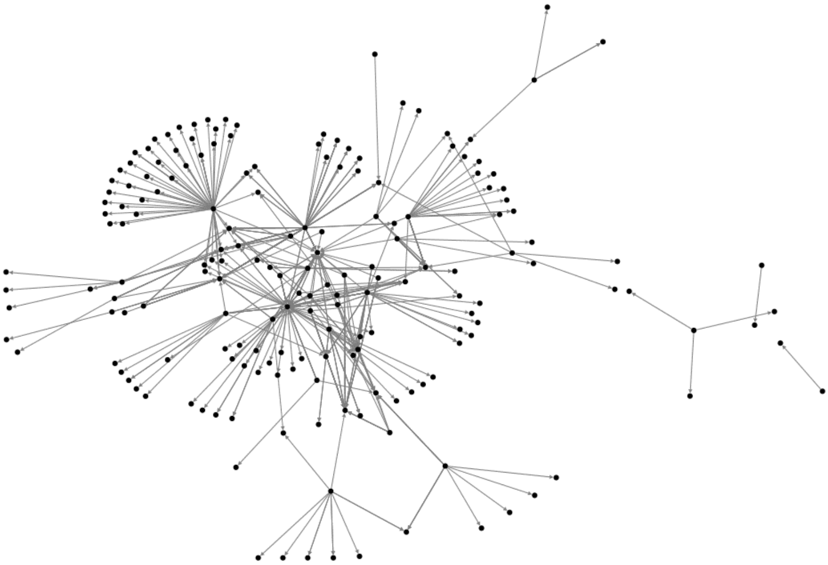

Figure 3

The whole structure of documentation system

Table 1

Overall metrics of documentation system

Metrics	Value
No. of Vertices	181
No. of Unique Edges	185
No. of Edges With Duplicates	300
No. of Total Edges	485
No. of Connected Components	4
Maximum Vertices in a Connected Component	173
Maximum Edges in a Connected Component	480
Maximum Geodesic Distance (Diameter)	8
Average Geodesic Distance	3,621674
Graph Density	0,008133824

Overall metrics mostly reflect the structural properties of the documentation network.

The rate of numbers of unique and duplicated arcs shows that elements of this system are complex documents and not small and task-related regulations units introduced in 2.4. In general, the bigger the rate of Numbers of Unique Edges to Numbers of Total Edges, the smaller part of the operation is regulated by nods of the documentation network. Obviously, it is true when only one type of connection is applied in the network.

The Number of Connected Components represents the connectivity of the network. Having 4 such components here means that this documentation structure is highly connected. This conclusion is supported by the fact that the far biggest part (173 nods) of all units (181 nods) belongs to the same subgroup.

The relatively high Maximum and Average Geodesic Distances 1 reflect that regulation chain is the typical structural element (see in 2.1).

The very small Graph Density 2 (8.1*10-3) denotes that references among documents are seldom.

According to the overall indicators mentioned above we can conclude that this documentation system consists of complex documents forming a highly connected network with regulation chains as a typical structural element and with relatively few links among the nods.

3.2 Individual Metrics

To identify the different roles of nods, individual metrics are calculated too. Their values can be seen in Table 2, Table 3 and Table 4. The associated subgraphs are highlighted in the following Figures.

The value of In-Degree represents the importance of a nod. The most important document – from this point of view – has 22 individual incoming links, – due to its 22 connected neighbors (see in Table 2 and Figure 4). We can realize its important role in the graph too, since it is located in the middle of the network. There are two more nods with more than 10 in-degree (14 and 11). For further reference, we call them ID1, ID2 and ID3.

Table 2

In-Degree metrics

Metrics	Value
Minimum In-Degree	0
Maximum In-Degree	22
Average In-Degree	1,464

[1] Geodesic distance is the distance between two vertices along the shortest path between them.

[2] Number of unique edges per maximum number of edges the graph would have if all the vertices were connected to each other.

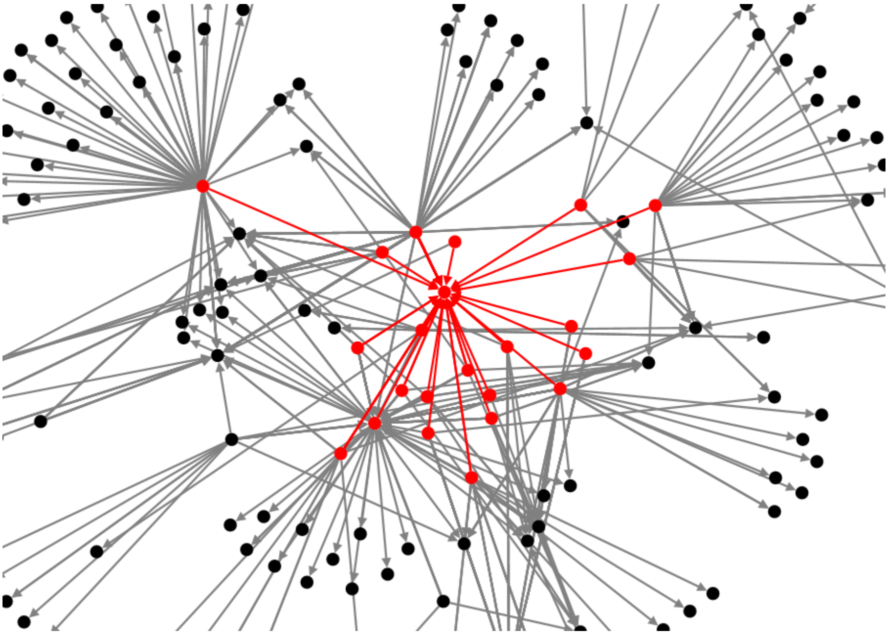

Figure 4

Subgraph of the nod with the biggest in-degree and its connected neighbors are marked in red

If we look at the list of nods with out-degrees, we can see that the most sensitive document (call it OD1) seems to have 44 outgoing links (Table 3). If we see the graph (Figure 5), we can realize that these are concurrent links, which means all of them connect only two nods. The conclusion is that despite the high value of out-degree, this document does not play a significant role in this network. It is interesting that out-degrees of D1 and D3 are zero, while D2 has the second biggest out-degree (24). However, D2 has only two neighbors, so its role is not significant either. Instead, the nod (call it OD3) with the third biggest (19) out-degree has 19 neighbors (Figure 6). It means that it is the most sensitive document in this system.

Table 3

Out-Degree metrics

Metrics	**Value**
Minimum Out-Degree	0
Maximum Out-Degree	44
Average Out-Degree	1,464

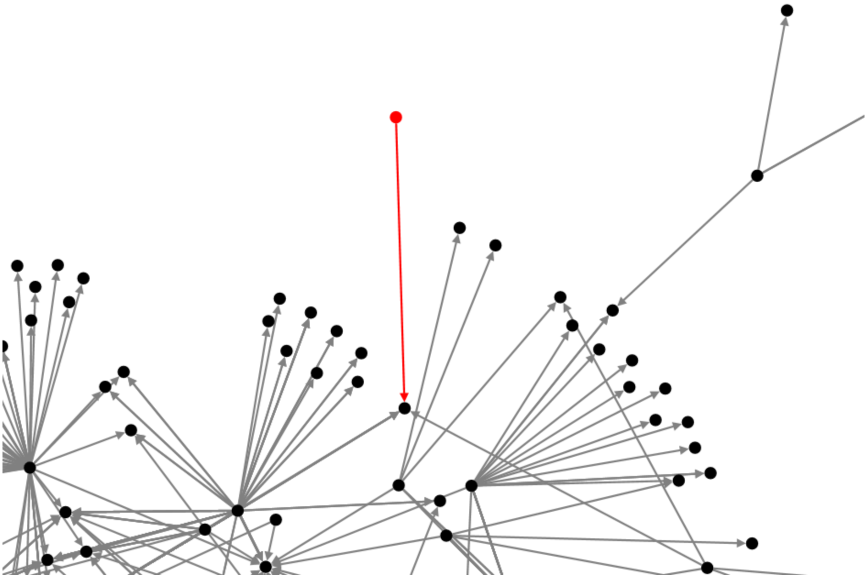

Figure 5

Subgraph of the nod with the biggest out-degree and its connected neighbors in red

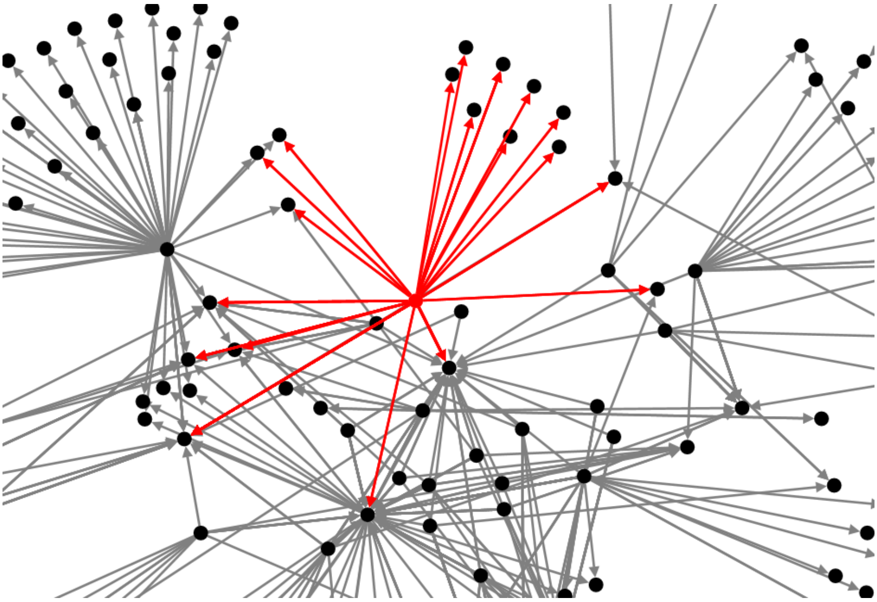

Figure 6

Subgraph of the nod with the third biggest out-degree and its connected neighbors in red

The averages of in- and out-degrees are equal. We could come to the conclusion that the sensitivity and the importance of the elements of this network are the same. If we check the degree distributions (Figure 7), we can see that there are very few nods with high degrees while most of the nods have much less connections. But we should not forget the fact that leaders of these lists have different numbers of neighbors, which influences the evaluation of the roles of the nods.

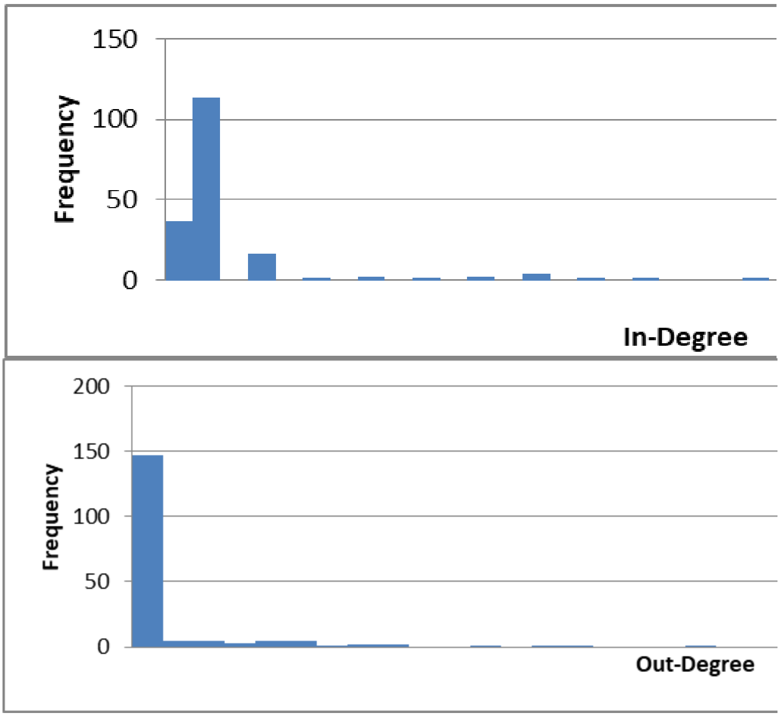

Figure 7

Degree distribution

As we wrote in chapter 2.3, Betweenness Centrality shows how critical the role of the nod is in connecting network parts. 5 of the top 6 nods in this list are the top 3 nods of in- and out-degree lists. The exception is the nod with the fourth biggest betweenness centrality (call it BC4) that has 14 out-degree and 0 in-degree (see in Figure 8). BC4 is a typical representative of network bridges, but in documentation system it is not so obvious. Here the different rates of in- and out-degrees show different types of bridges. If it has few in-degrees and a big number of out-degrees, the nod is a so-called fork-bridge, which means it is a rather sensitive document. On the other hand, a nod with few out-degrees and a big number of in-degrees is a join-bridge, which means it is an important document. If any of the degrees is null, the nod is not a real bridge.

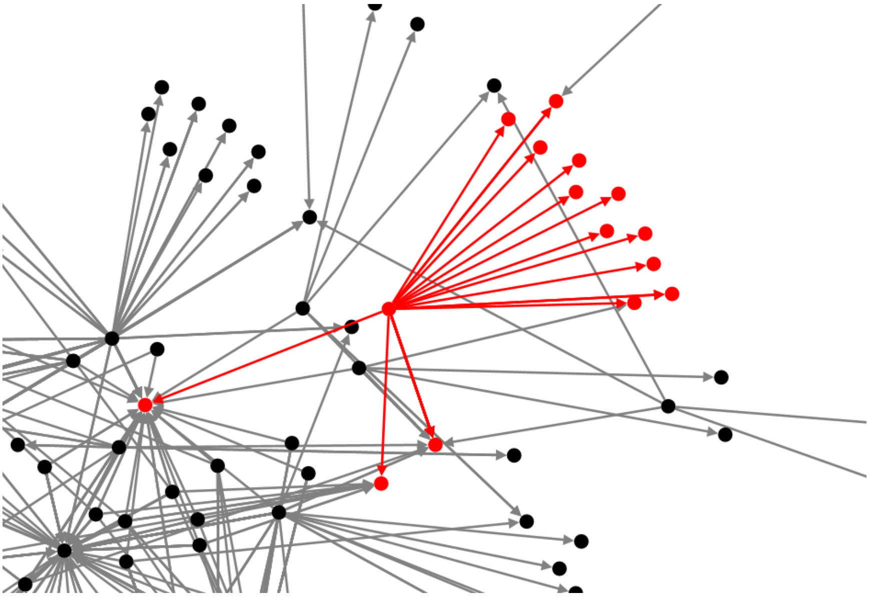

Figure 8

Subgraph of nod with fourth biggest betweenness centrality and its connected neighbors in red

The Closeness Centrality is the reciprocal of farness, which reflects how far a nod is from other nods it is connected with, i.e. in network terms how long the paths are between the connected nods. In documentation networks the smallest closeness centrality belongs to the documents that take place in long regulation chains. In our sample network 173 nods have 0.001, 0.002 or 0.003 value, and only 8 nods have more (0.2 for 3 nods, 0.333 for 1 nod, 1 for 4 nods). It means that there are typically long regulation chains between documents and most of the documents take part in these paths. Documents with high values are the part of regulation islands, i.e. isolated groups of nods. Such distribution of closeness centrality shows that this documentation system is uniform but has a very long and complex set of connections that makes it difficult to easily overview and understand it.

Another type of centrality related indicator is calculated for our sample network to highlight the importance of documents more sophisticatedly. This is the Eigenvector Centrality, which takes into account not only the number of connected nods but the degree of connected nods as well. It means that in documentation networks the bigger degrees the neighbors of a selected document have, the more important or sensitive the it is. The distribution of values (Figure 9) may be more interesting than its nominal value (Table 4). It demonstrates that there are noticeable differences among the eigenvector centrality values of nods. In our sample network it fine tunes the description of importance of documents.

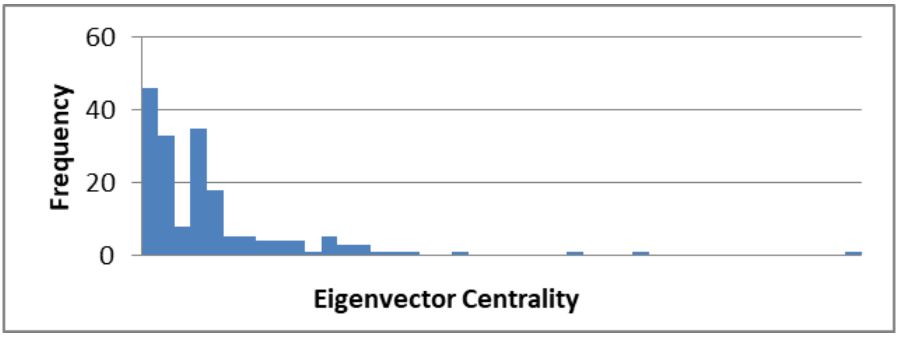

Figure 9
Distribution of eigenvector centrality

Table 4
Eigenvector Centrality metrics

Metrics	Value
Minimum Eigenvector Centrality	0,000
Maximum Eigenvector Centrality	0,057
Average Eigenvector Centrality	0,006

Conclusions

DQNET can be applied to identify and map quality regulation networks, to describe the properties of documents and to identify optimization opportunities. To conduct these activities properly individual and group network indicators have to be reinterpreted according to the specific characteristics of the documentation system. As an example, hubs can be important or sensitive documents, bridges can be fork- or join-bridges, subgroups of nods can be regulation-chains of regulation loops.

References

[1] Cao, Z., Wei, F., Dong, L., Li, S., Zhou, M. (2015): Ranking with Recursive Neural Networks and Its Application to Multi-Document Summarization, pp. 2153-2159, Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence

[2] Carley, K. M. (2015): Crisis Mapping: Big Data from a Dynamic Network Analytic Perspective, Journal of Organization Design

[3] Tan, L. S. L., Chan, A. H., Zheng, T. (2015): Latent quality models for document networks, arXiv:1502.07190v1, Annals of Applied Statistics

[4] Wan, X., Yang, J. (2008): Multi-Document Summarization Using Cluster-Based Link Analysis, pp. 299-306, Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval

[5] Wang, C. (2015): KnowSim: A Document Similarity Measure on Structured Heterogeneous Information Networks, Data Mining (ICDM), 2015 IEEE International Conference on

[6] Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., Hovy, E. (2016): Hierarchical Attention Networks for Document Classification, pp. 1480-1489, Proceedings of NAACL-HLT

[7] He, Y., Wang, C., Jian, C. (2017): Modeling Document Networks with Tree-Averaged Copula Regularization, pp. 691-699, Proceedings of the Tenth ACM International Conference on Web Search and Data Mining