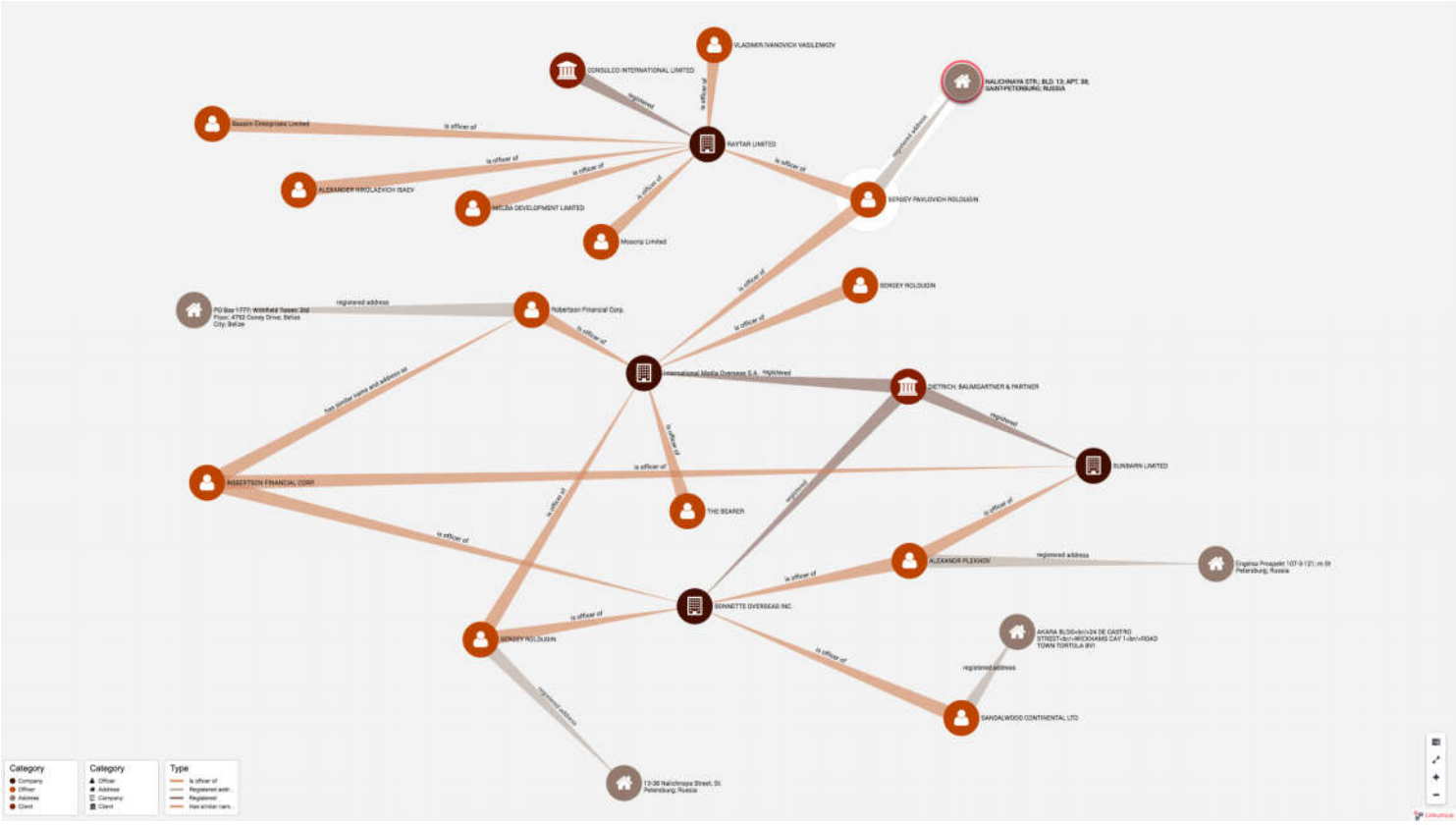
Technical aspects of "Panama document" analysis



Company and officials of the network that appears to be involved in the assets concealment of President Vladimir Putin (Generated With Linkurious)

It has become a hot topic all over the world Panama document . It also becomes a situation, such as shaking the regime in each country, but when viewed as a purely data, I think this is interesting also to those involved in computer and data analysis. For such data of contents and background have is terrible news because it does not go here. On the other hand, what kind of work in the field has been done is not less coverage. As a practical matter, if

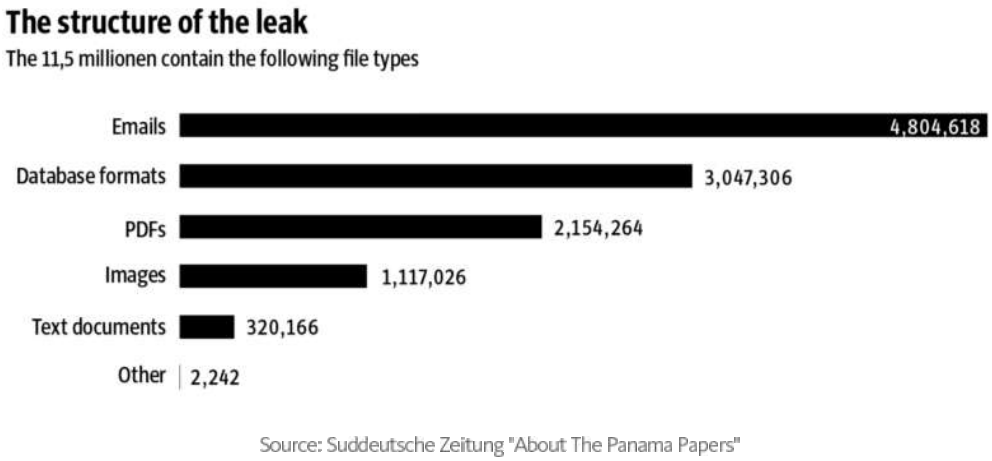
you got a leak data of helpless amount in the manual, do you research news organizations have done what? I also I was wondering from an earlier, the other day database companies that are and, company to create a data analysis application in the blog, there was a post that can be glanced the actual one end:

- Panama Papers: How Linkurious enables ICIJ to investigate the massive Mossack Fonseca leaks
- The Panama Papers: Why It Could not Have Happened Ten Years Ago
- Inside the Panama Papers: How Cloud Analytics Made It All Possible

Since these were also interesting for me that is the job of dealing with the graph (see below) from the usual, we will introduce here. Since these technical part of modern investigative journalism will be are rarely reported in the Japanese media, it was purely interesting I was reading. Here, investigative reporting agency of the world will know what you are doing to use any technology, even in people not bright in technology so I'll write as much as possible easy to understand.

The format of the data

First is the computational nature of this data. According to information published, it is something in the manner described below



- Capacity: 2.6TB. The size will fit all ten thousand yen of the hard disk.
- Number of files: approximately 11.5 million

- Data format: e-mail, databases, such as RDB, PDF documents, image (perhaps many scan documents), text file. Refer to the chart above distribution of the number of files

Data that 2.6TB is, never is a big thing for the current computer. You If you have a HD recorder for TV recording, probably can be stored there in a big data than this. However, when I am going to analyze the data of the size humans, it is in the amount also enormous too, and is not meant to be somehow a very human power. For essentially this data set, anachronism approach, such as that seen every single print out will not work. **If Panama document is not available some sort of search and visualization of technology, but human beings of such hard to chew things to.**

However Fortunately, if this time of the data except for the image file (also recently "intelligent" analysis technology has dramatically improved, but does not mention here), a relatively low cost almost entirely by machine in the file of the process can form. Here it will be to turn the computer. From this point on, I want to explain about the technical background that can be read from the previous article.

Pre-processing of data

If you want to analyze such a huge number of files, you will always need a pre-processing of data. This time of the data set can be roughly divided into two types:

1. **Easily accessible files in a machine of RDB format (so-called database)**
2. **Document file on the premise that the human-readable. Including text and images, the PDF**

Data analysis team of ICIJ that hit in this analysis, first worked in a relatively hurdle low one day of data. In short, it is to re-build in a form that can be easily search the database. This is the case was able to finish in a few months at the hands of the experts. But second of data does not go so.

Conversion to text data

This article According to, many of the image seems was something you scan documents of paper, if an attempt to get your character information from

there, technology called OCR is required. OCR and is a technology that is cut out to recognize the characters in short from the image. A lot of users in Japan Evernote is a technology that has been used in such. Character of the label of the drink, which was reflected in the photograph I think that may be caught in the string search, but things that technology. Can be carried out by a combination of laptop and home scanner recently, is widespread technique. However, since the number of this time image is enormous, in order to shorten the time, which is a commercial cloud computing services (in short of the computer time is a rental shop) Amazon EC2 of Web Service seems to use. What instance type to do, but could not be confirmed, using 40 units of EC2 instance from 30 units, we first a single-mindedly work to convert the image data to the character information.

Graphing Data

When you do the text of by OCR, the treatment will be easy in the time being computer. Speaking extreme that, as long as you have saved in a text, even by string search open it in the editor, you will be able to read or is written is what is to some extent. However, the detailed investigation there and eventually scanned documents and financial reports must be read by experts, from the people and companies of the network, it must to be examined in more detail in the keyword search . Is what that used to build it called full-text search engine. What was actually used Apache Solr in, the cut-out of the meta data and text Tika was used is. Utilizing these, they were first cut out the metadata from the set of files. Here, the meta data to say, in such as file type and time stamp, will be easier to search a large number of files by indexing it.

However, the present data are approximately 215,000 objects company (perhaps many paper company) is included, another twist is required in order to analyze the money flowing therethrough. If you are reading the accounting report and registry of a company, the company of officials, and if to be seen the connection of such other companies associated with the company, a great help to understand the flow of money you. Of course, this can be expressed in the text. For example, experts will that you have obtained the following information by reading the current document actually:

- Mr. A's current president x
- A company was founded by Mr. Y is a director of Company B
- Mr. Y is sending frequent e-mail to hit a large amount of the keyword "Company A" to Z Mr.

- There are A company to address that α, B's location is a β
- The α and β of Grand Cayman Island, exist on the same floor in the γ building
- Mr. Z is the owner of the γ building

If this level of connection, from here

> *The Mr. X and Mr. Y there is a connection that A's founder and its successor, B Company there is some sort of connection with the A's. And although there is no name on the registration of a company, in the Z and Mr. A's probably there is some relationship through Mr. Y. And from the two companies of location, these are there is a possibility of a paper company in which the same broker was established involved. The broker there is a possibility of Mr. Z.*

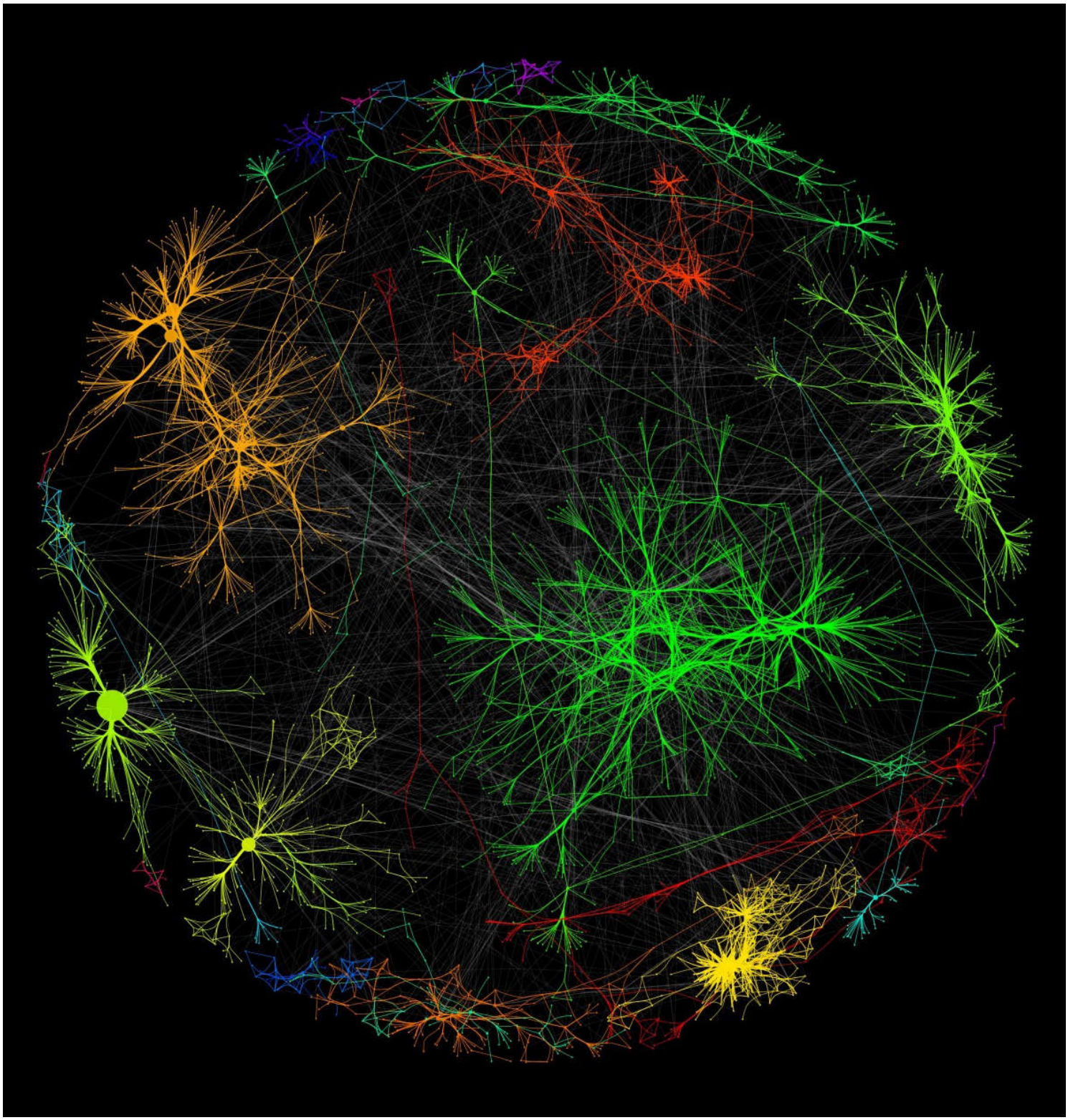
Reasoning, such as that it is also possible to human beings. But what about such a link hundreds of thousands, are present at the level of millions. It is impossible to grasp one by one no longer writing down the sentence the whole picture to human like this. They are for the **graph of the data** was selected.

What is a graph?

Hear the word "graph", such as a line graph or bar, I think that many people think of the chart, which is called in English the so-called "diagram" or "chart". But graph referred to here is the graph of mathematical terms. Strictly speaking, it will be defined as follows:

> *Graph G is the ordered pair consisting of a set E of the contact set V and the sides of, G = (V, E) and expressed*

Is a small difficult, but what a thing is not just a nickname in the world of mathematics of just a "connection", or "network". Seeing is believing at first glance, I think that it is obvious if you look at the following figures.



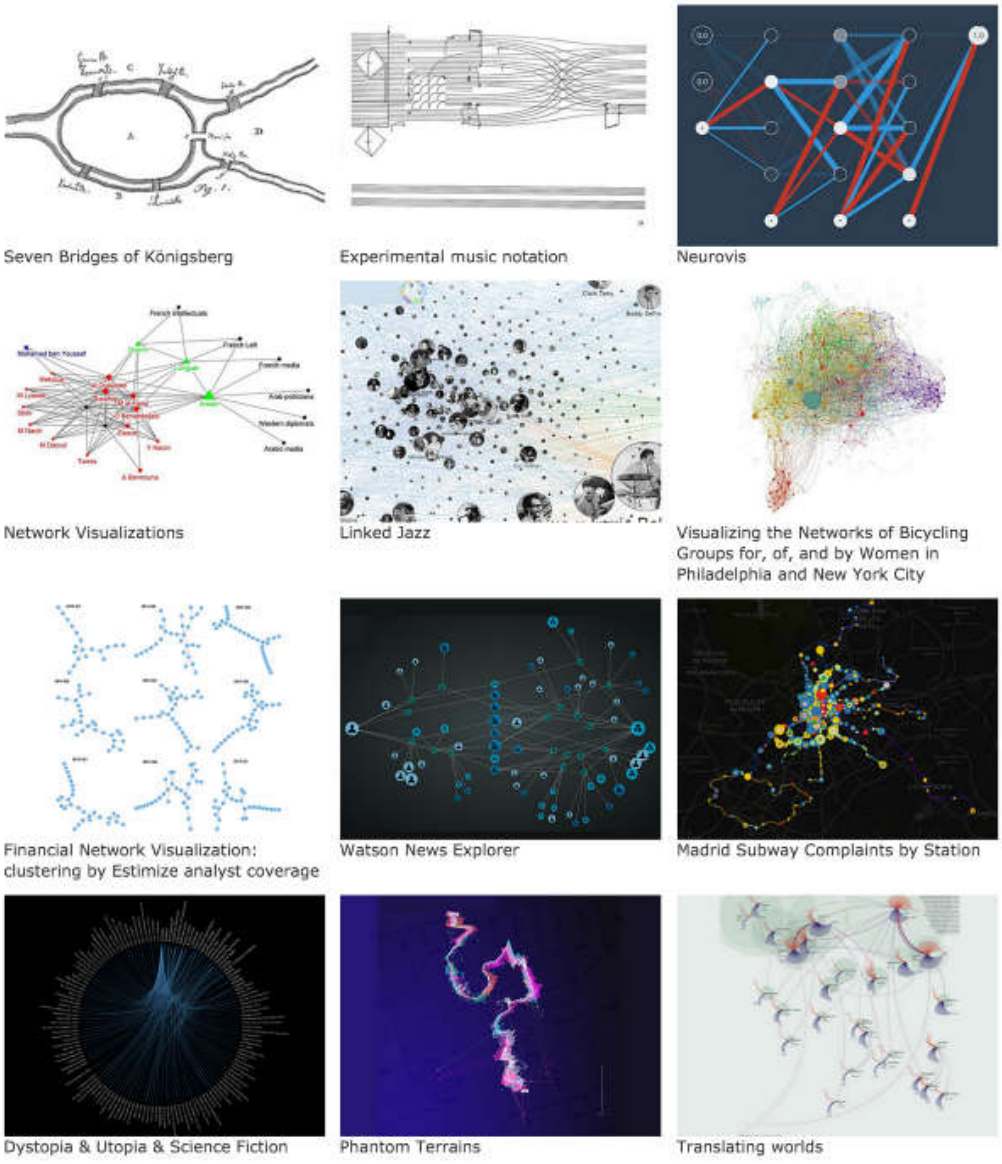
Protein-Protein interactions of yeast. Generated with R, igraph, and Cytoscape. By K. Ono. CC BY 4.0

That such a network structure ***graph*** is called. Each vertex in the example above (***node***) of yeast to use when baking bread protein present in the, side (***edge***) has means that they cause a direct interaction. Even in the biological

field of my involved, this way or represents the nature of the relationship between the substance, in order to the biochemical reaction in the extremely complex body person is easier to understand, those of express the reaction path in the graph to approach it said that it is used frequently. In other words, any route and, relevance of the person and the person, if you want to analyze things like interaction, it is very important that the data is computer be expressed as a graph of the form that can be chewing, to human beings by performing it you will be able to analyze the level not possible.

Thing called the structure of the graph, you literally appear everywhere:

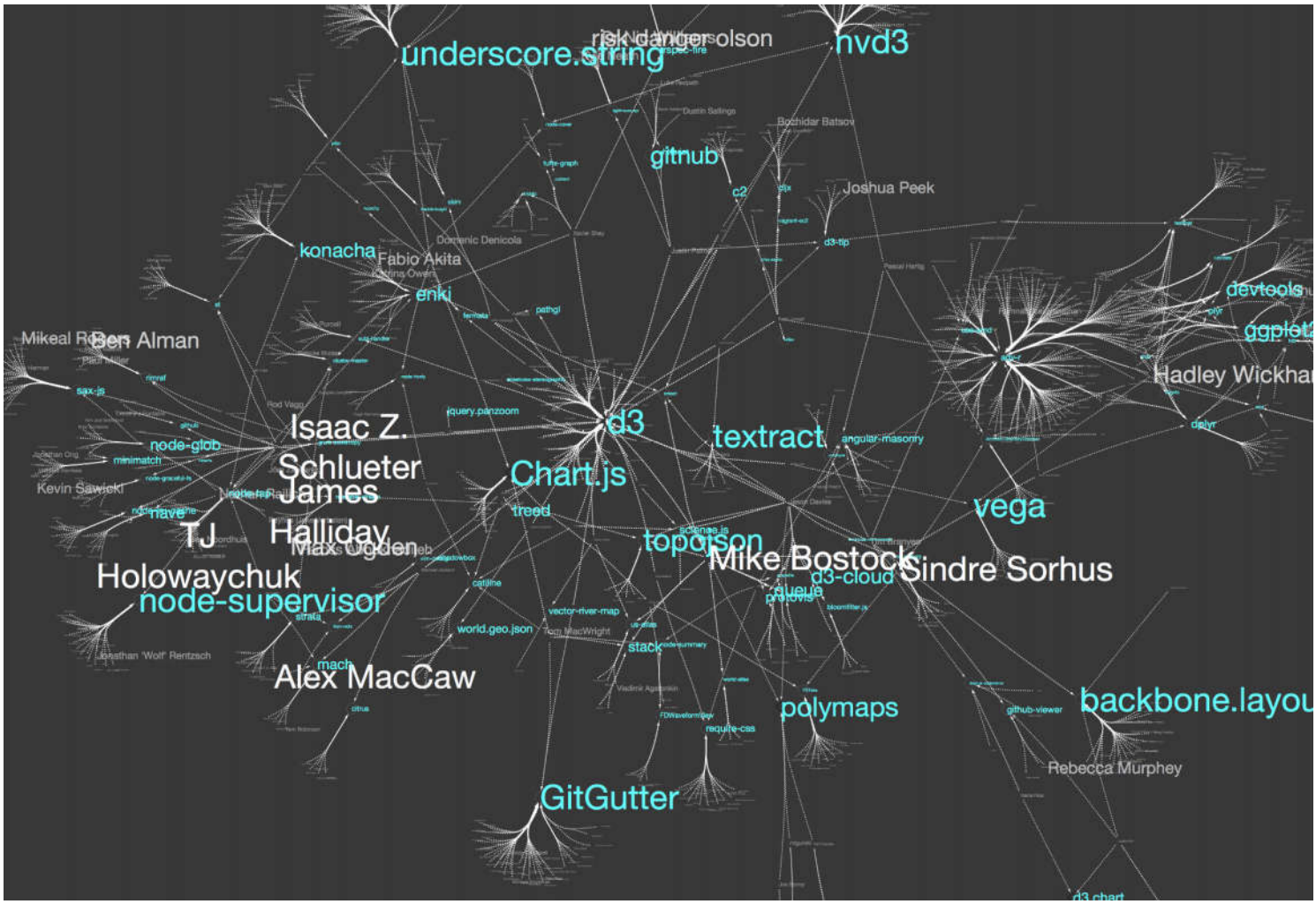
- Social networks
- Protein-protein interaction
- Expanding the route of infection
- map
- Organization within the structure of the criminal organization
- Romantic relationship



Complexity Visual : long-established sites that collect the graph visualization of all types

And that to appear such a graph to the human eye ***graph drawing*** is called. Because the structure of the graph appearing in a variety of fields, itself be drawn or to parse this structure has become a field of study. In recent years, with the spread of such a social network service, because it is easier to obtain data in complex networks in large, various studies have been made in the field of social sciences. By the spread of this kind of research, graph analysis and visualization can be used by anyone in the open source, software that can be used for accumulation of data was spread rapidly. But typical, the old graph with a history drawing software of ***graphviz*** , ***R*** runs on statistical analysis for the programming language execution environment to be used in favor of the experts of the statistics that ***igraph*** developed, software was also used in this

case to have the person was involved ***Gephi*** , is a database in order to be able to search and store data of large-scale graph ***Neo4j*** , also used to create the diagram some of I made to come out to this article are, our team has developed ***Cytoscape*** and so on. For almost all of those can be used for free, is the current situation is that of advanced graph analysis and visualization technology is becoming more common.

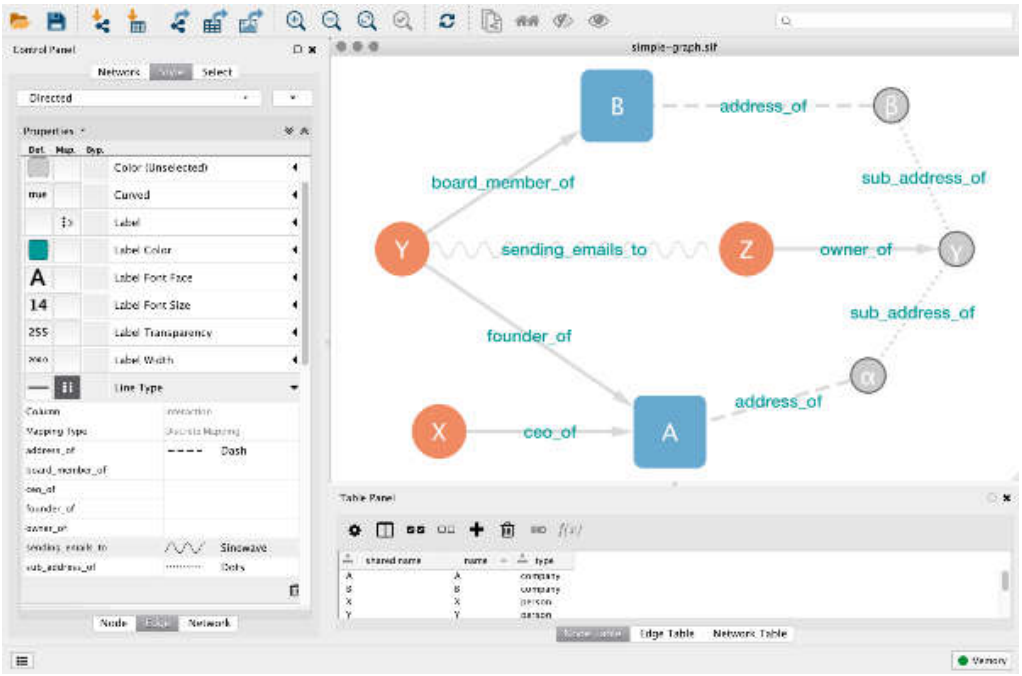


Visualization of Dataviz community that exists in the GitHub By K. Ono. CC BY 4.0

Actual graph construction

"**To build a machine-readable graph data**" I very pompous and write, but actually not too complicated. Basically, every edge can be expressed in the following format:

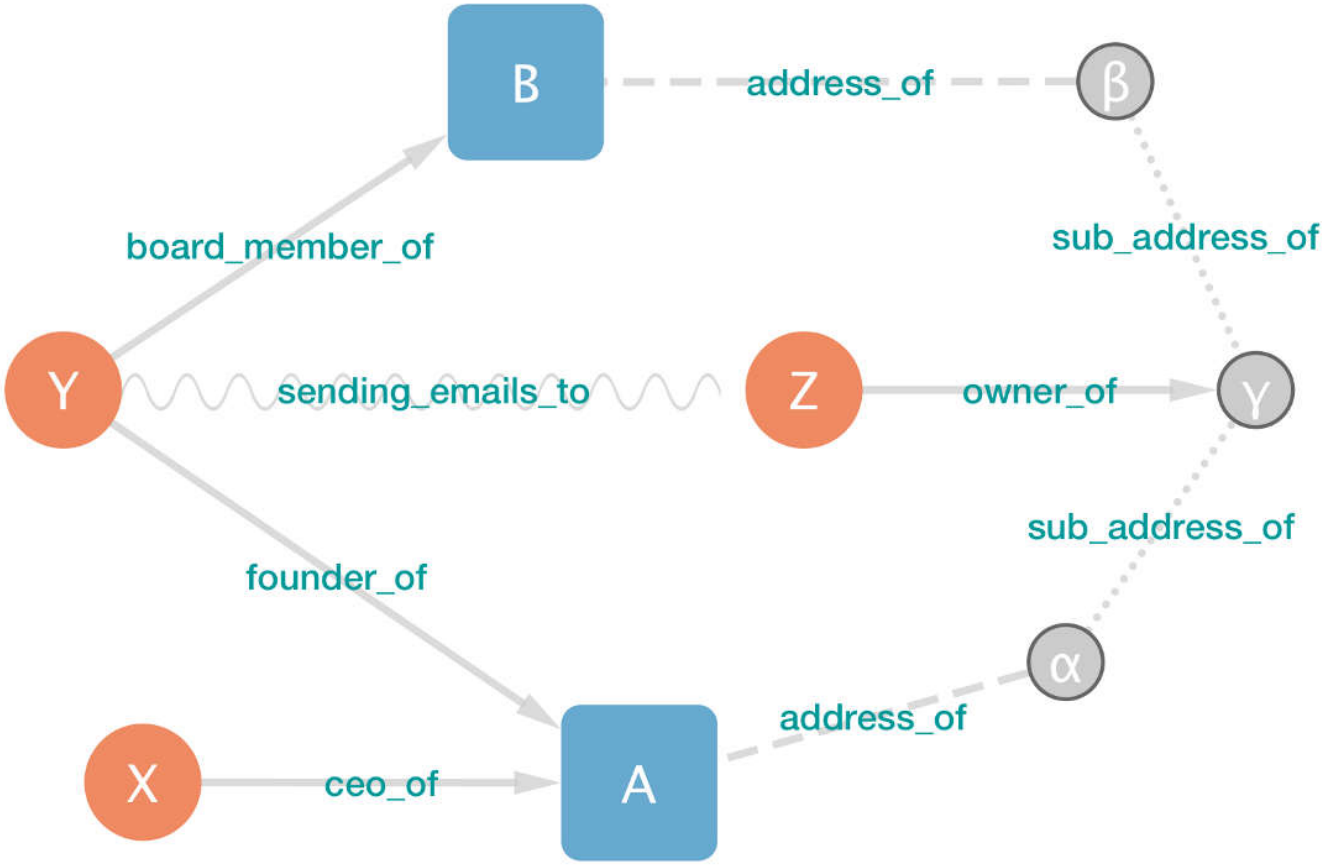
> *Node A, Edge Type, Node B*



This means that if you can describe the relationship between the text of such a form, every graph can be used as data used by the database and visualization software. More specifically, is the graph of these often also be published as a text table of mere CSV format. In fact, those in the case of offshore Leekes incident is a spill incident prior to Panama document, graph data in this format has been around. In order for you to know how much simple Let's actually made. The data we use the example of a fictitious data that was used as an example in the previous section "suspicious gentlemen of the Cayman Islands". And to express that sentence as the graph will look like the following:

> ***Ceo_of A X***
> ***Y Founder_of A***
> ***Y Board_member_of B***
> ***Y Sending_emails_to Z***
> ***alpha Address_of A***
> ***beta Address_of B***
> ***alpha Sub_address_of gamma***
> ***beta Sub_address_of gamma***
> ***Z Owner_of gamma***

And save it as a text file, Cytoscape something like this can be seen when you read visualized



I think many people was that the are made by hand something like this in such a complex plot mystery novel of the eyes. Essentially, but's the same, is a big difference, **mechanically if Totonoere a mechanism to create such data, to draw the search or relationship to the machine with the keyword even a tremendous amount of data human There can be to easy to shape understanding** , is that. Not mentioned is because the technique of the details out with this purpose, to try out such a visualization, you can easily open source software, such as a text editor and Cytoscape. Data using this time is set aside here:

- Sample Cytoscape Session File

So, what would the graph of this data was what kind of work. Because it does not have the actual data is public, but mingle also expect more information on this part, there are some hints. It is a very small part graph, but this article embedded viewer of Linkurious that can be manipulated interactively has been published.

"How Gunnlaugsson hides his secret assets" (correlation diagram around Prime Minister resigned Iceland)

As long as you see this, the node Yes classified into the following types:

- Officer
- Street address
- Company
- Consulting companies

And it "an officer of the company" are classified into a relatively simple relationship, such as "registration destination address" we are building the

edge. From here, you probably imagine that or not is was going to be such as the following:

- First, the company is a node from such registration, connection attendant officer, address, a consulting company, which is in charge as the related node
- Performs a name identification of the node, grant the edge as there is relevance in between, such as a similar company name, person's name
- This single-mindedly repeat, continue to rapidly connect the graph

If the sub-graph is intended nothing much different from the original database that have been published if, I feel that together as a relatively simple data. In this case, there is also a possibility that Horiokoseru future relationships that are still asleep. For example, since it has this time also includes the e-mail data, to measure the direction and frequency of communication from those of the sender (From) and destination (To), converted from there to score the edge. Since this by the edge of the weighted occurs in large quantities, I think it is also possible to finish in a more "high resolution" graph data. Of course to its destination, but there will be a variety of analysis using the technology of natural language processing, it does not go this time. It should be noted that the technique of estimating the network structure of the relevant person from the e-mail data, and the like have been used in various situations, the United States of the intelligence agencies of the terrorist network in order to or visualization analysis, are using a similar approach it is said that. As an aside, in the past a huge scandal Enron from mail archives that are published as evidence in, there is also a project that the network of officials say that to visualize. The author then Stanford, become a professor over the University of Washington, from his lab D3.js was born library for a very famous visualization that.

- Enron Exploring: Visual Data Mining Of E-Mail By Jeffrey Heer

Well, after you decide a policy to make a graph, intently processing and cleaning of data, it is poured into the database. Have you started working in the world of data analysis I think you can see, but this is also a very painful work. So-called "data science" is actually why it is said that it is a very gory work is also available around this. In the present case of the data, to the company organization of 215,000, cage least three parties, it was the graph of

about a million nodes and graphing these ties is finished. When you actually build a database of Neo4j is like using this software

http://www.talend.com/download/talend-open-studio

I do not know and still do not do interviews towards the person in charge of detail, generally in such work, it seems graph database that first the investigation of the backbone has been constructed.

Read up to here, for if the people of the RDB of experience who had leaked the data in the file of "the first place RDB, it's throw a SQL query to the database that was why integrating this what do that Madorokkoshii? As it is is not it? "and it might be no doubt. The biggest reason that they were purposely graphed, if you do a search specific to the graph structure, such as the route search in large quantities, I think it's because performance much not to the use object's what was built with RDB is reduced in some cases . In this study, because it was the use case that read the article while looking at the connection of the related person with the flow of gold, Will to use a graph DB was determined to be the most suitable.

Human wave Linkurious for tactics

It will be a little earthy from here. Era, such as "to automatically analyze what was a database to artificial intelligence" does not come at least still for the time being. So, it is a man of the turn from here. The first place to the big reason they have to use a graph database, the connection of the parties as human beings literally can be seen, it seems that a large part that I wanted to use the functions of the graph visualization that. If you want to understand the complex graph structure, even if not read the relationship in a sentence, so the human brain can not be grasped in a short period of time the large ones. So they the Neo4j database built Linkurious were connected to graph visualization and analysis service called. In this analysis are involved journalists who scattered to 370 people around the world. If these people are going to go the analysis work and dispersed to access the same data at the same time, we need to build some kind of application that allows such a work on top of the graph database. Of course you can also create their own custom applications, but to do not need to do up there, it is because too large burden, seems to have to use a service called Linkurious Enterprise to provide the Linkurious company.

The majority of this time participating journalists, using the GUI of the application that Linkurious-supplied, pull out the relationship between the person and the company from the graph database, it seems to read the relevant documents and materials while checking it in the eye. Of language for graph queries that are standard on Neo4j during Cypher seems Moi users with advanced technology, which was analyzed while a more advanced search in. Because graph query language is, to a certain graph, it is a kind of DSL (domain-specific language) of the order to search a route or node that matches the specific conditions. It is generally correct even if asked to think like a simple programming language to search for the graph. Speaking in a simple example,

"Whether enumerate the people and companies that are connected within two hops to Company A,"
"examine whether there is a path between the company and Mr. X with a certain country president and involved. If there is case enumerate it all."

The search condition of feeling, is a language to replace the search criteria in the form of a computer can understand. One of the advantages of using a graph database in these investigations is that it is possible to perform fairly complex queries to the structure. Route search, as well as with or find a part that has a specific small structure (also referred to as the network motif), can be done in a realistic time in RDB difficult analysis. Perhaps because the still supposed intended to make also a variety of analysis the future, I just want to wait for further news.

Summary of the technical points

- Do the reverse engineering to RDB, it was to be extracted to find an integrated form schema
- Use an instance for OCR, which was prepared on the AWS, it was the text of a large number of files
- Cut out with Tika the meta-data of the file, poured a variety of documents that have been text into Apache Solr of full-text search engine. Thereby it was to be able to keyword search
- Data was processed using the Talend, and graphed, it was stored in the database Neo4j

- It was adopted Linkurious Enterprise as the front end of Neo4j. As a result, the user who distributed can now access from the GUI at the same time complex data sets around the world

Forecast of future

Is still (raw) because the data is not out of things can not be said only in the imagination pains, the impression that I read these articles, by to welcome the people who called the so-called data scientist, it can be done still multi-faceted analysis I feel that so.

Kind of edge that has come out in the sub-graph of Linkurious that is currently being published has been quite limited to, it will be visible to as not doing advanced things, such as the flow of the mapping of the funds of the scoring and time series. It is not a statistical shop, but only is the opinion of the position as a programmer. To collaborate and share data in secret, but is Linkurious that played a major role as a front-end in terms of, because Neo4j of the back-end is to exist independently, analysis and graph structure itself future, edge / If you want to further multi-faceted analysis of the data by the property added, it is likely that to consider the complex scenarios via direct Cypher. Moreover, keywords or extracted from unstructured data, increasing the thickness of the graph database like scoring edge made based on certain rules, expert with advanced analysis capabilities Jupyter Notebook reproduced open using such do the sex of some analysis, notes and data, summarizes the environment to Dockerfile hopefully, it will be published as it is, that I think that there may be road. Will if so from the analysis of the environmental approach, in a sense the ultimate of public information, including to data.

As more of the person in charge is he said, What is interesting is relatively easy Swiss Leekes is the integration of the data set of. Since they began to use a graph database such from the time of the single item of Swiss Leekes, if there is something that has already been graph database, merging it it seems that it is relatively easy. As a result I think that interesting if it is even larger overhead view. She also You have said this, let's wait for the follow-up.

> *I think that we have just scratched the surface on how we can analyze the graph data.*

> *We are for terms of how can analysis the graph, is the stage of the degree to which still scratch on the surface of the problem.*

> *Mar Cabra, ICIJ Data and Research Unit Editor*

Also, all data public is unreasonable, if it is published, even partially as to some extent processed secondary easy-to-use data, infographics and data visualization team, such as newspaper, various types of public data sets (such as map) merge to go with, it might be possible to present a more easy-to-understand the whole picture.

in conclusion

What did you think. By the way tool listed here, except for the Linkurious, I made a sample of visualization Cytoscape available at all also including Neo4j free. The data structure that chart because it appears everywhere in this world, be stored it in the form that can be read by machine, I hope you know and in this way can be used in various situations.

Surprisingly, this time of work is all organization of journalists we are. They have a data analysis group therein, the men from the pipeline created for the pre-processing of data, data processing in a manner that does not have a computer at their own using a commercial cloud services (perhaps many on AWS spot instances use batch processing by), system architecture for distributed work, construction of the database, you're doing such as UI of choice, things like conduct in the software system of the company on their own. Tip of the world of investigative journalism was impressed with what is coming up here. At the same time, I felt also or not is interesting is also such thing as a career path of software engineers.

But is today a keyword, such as grab the big data Dano what Dano clouds flying, here is spelled out in one carefully Jupyter Notebook data analysis pipeline to combine your favorite tool, analysis and visualization of an impact on something society Why do not you also in the work? Now the work of the level that has not been considered in the old days I mean to acquire the skills and knowledge, using the open-source software and cloud services, is the era that enables even individuals and small organizations. This is some people referred to as the democratization of technology.

Finally we have quoted the words of Neo4j's CEO.

> *The democratization of technologies to make sense of data at scale is an important part of a free and open society, and I'm proud of the role we play in that evolving landscape - not only in the case of Swiss Leaks and the Panama Papers, but in solving future problems we can not even yet imagine.*

> *Democratization of technology in order to understand the large-scale data is an important part of a free and open society. And I'm proud of the role we have Ninae in the scene to continue its development. Not limited to Panama documents and Swiss Leekes, we still have to resolve the problem of the future can not even imagine.*

> *Eifrem Emil , CEO, Neo Technology*

4/9/2016 Keiichiro Ono

. . .

4/10/2016 postscript

After this, the Neo4j's blog, was published articles mentioned for more technical part. In fact it does become what I feel and do the construction and modification of the graph using the Cypher, because it is written with the code, please by all means If you are further interested in the details of the graph database. .

- ***Analyzing the Panama Papers with Neo4j: Data Models, Queries & More***

Any questions or problems, please e-mail to kono at ucsd edu.