Network Analysis and Web Analytics

THE CONTRACTOR IS ACTING UNDER A FRAMEWORK CONTRACT CONCLUDED WITH THE COMMISSION

Eurostat

Outline

1. Introduction to Network Analysis and Web Analytics
    a. OSINT: Open Source INTelligence
    b. Infrastructure Review
    c. Technical Skills
2. Information Sources, Retrieval and Extraction
3. Network Analysis
    a. Graph Construction
    b. Connectivity Degree, Connected Components and Giant Component
    c. Centrality Measures: Influence
    d. Communities
4. Practical examples using R and Gephi

Outline

1. **Introduction to Network Analysis and Web Analytics**
    a. OSINT: Open Source INTelligence
    b. Infrastructure Review
    c. Technical Skills
2. Information Sources, Retrieval and Extraction
3. Network Analysis
    a. Graph Construction
    b. Connectivity Degree, Connected Components and Giant Component
    c. Centrality Measures: Influence
    d. Communities
4. Practical examples using R and Gephi

Introduction
- OSINT

Open-Source Intelligence (OSINT) is defined as **"produced from publicly available information that is collected, exploited, and disseminated in a timely manner to an appropriate audience for the purpose of addressing a specific intelligence requirement"**

in Sec. 931 of Public Law 109-164 "National Defense Authorization Act for Fiscal Year 2006"

http://www.oss.net/dynamaster/file_archive/040320/fb893cded51d5ff6145f06c39a3d5094/OSS1997-02-33.pdf

Introduction
- OSINT

DATA **INFORMATION** **KNOWLEDGE** **INTELLIGENCE**

Tools and Methods

Human Interpretation

Strategic Application

European Commission

Intelligence Cycle

The Intelligence Cycle is the process of developing raw information into finished intelligence for policymakers to use in decision making and action.

Commonly, there are four to six steps in the intelligence cycle.

Introduction - OSINT

Federal Bureau of Investigation (FBI)

FBI is an intelligence-driven and threat-focused national security organization with both intelligence and law enforcement responsibilities in the United States.

FBI intelligence cycle describes the process of developing unrefined data into polished intelligence for the use of policymakers. They define 6 steps: Requirements, Planning and Direction, Collection, Processing and Exploitation, Analysis and Production and Dissemination.

https://www.fbi.gov/about-us/intelligence/intelligence-cycle

Introduction
- OSINT

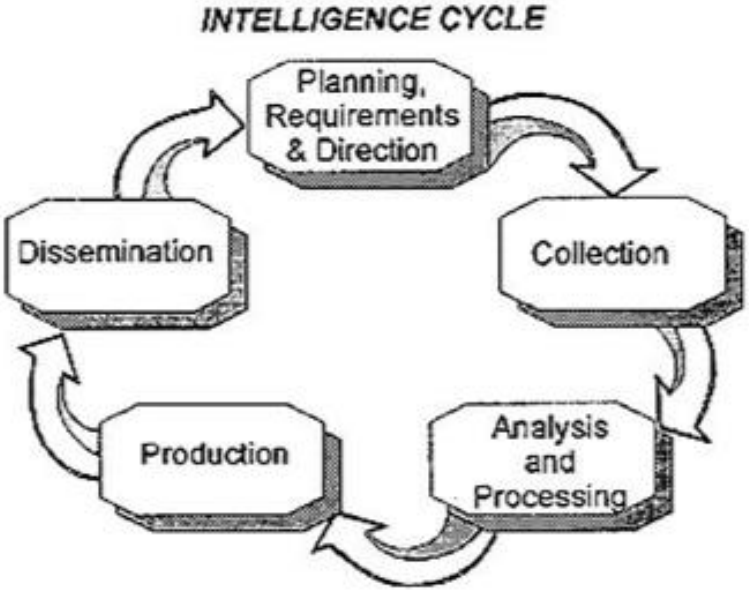
Center for Investigation and National Security of Mexico (CISEN)

CISEN generates strategic and operational intelligence with the aim at preserving the integrity, stability and permanence of the Mexican State.

CISEN defines a five-step cycle with: Planning, Gathering, Process and Analysis, Dissemination and Exploitation and Feedback.

http://www.cisen.gob.mx/intCicloInt.html

Introduction
- OSINT



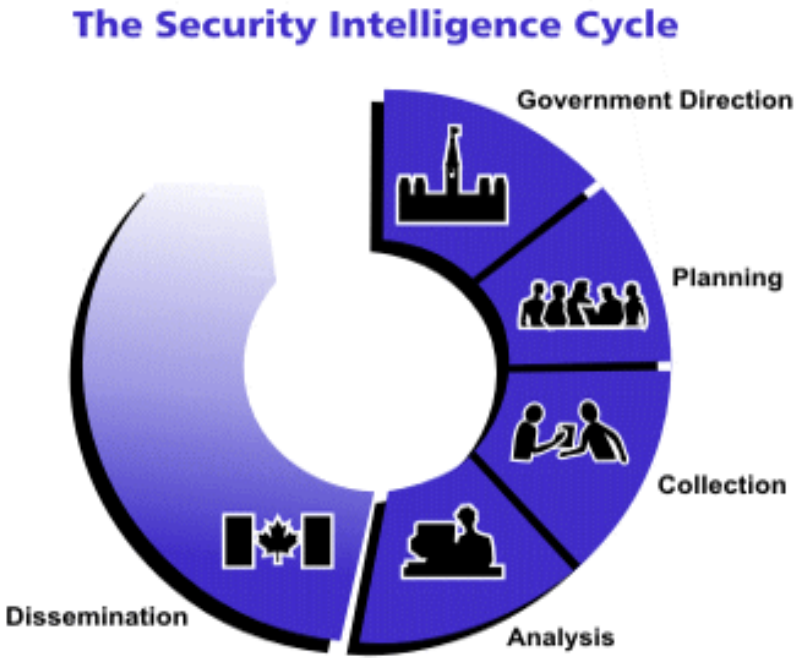
Central Intelligence Agency (CIA)

CIA acts as the principal adviser to the President of the United States for intelligence matters related to the national security.

CIA gathers, prepares information and generate intelligence reports to help United States' leaders to make decisions. It defines a cycle with five steps:

https://www.cia.gov/kids-page/6-12th-grade/who-we-are-what-we-do/the-intelligence-cycle.html

Introduction
- OSINT



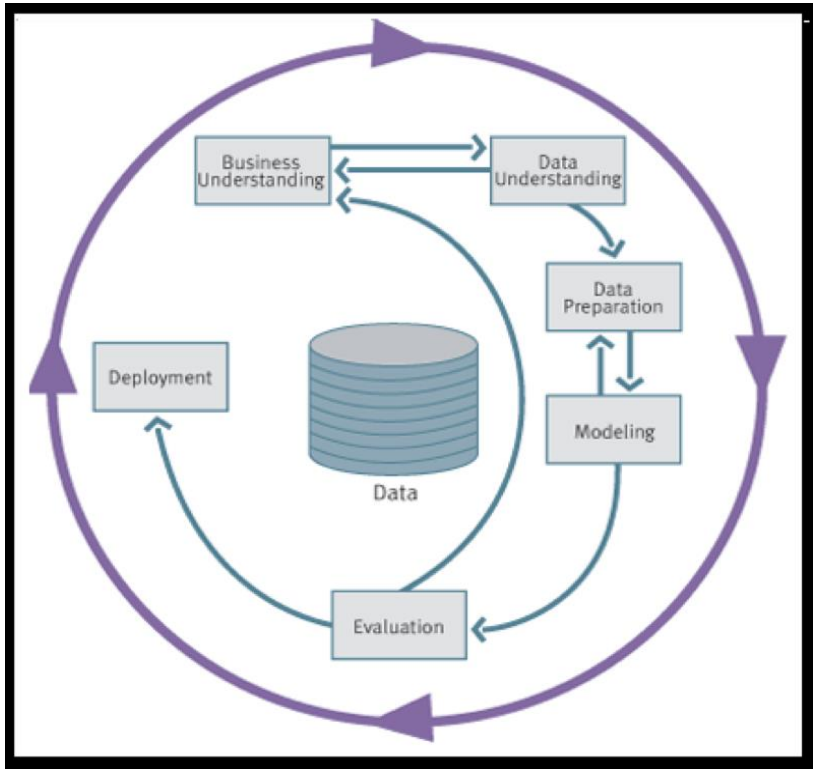
Canadian Security Intelligence Service (CSIS)

CSIS is at the forefront of Canada's national security establishment. Its main role is to investigate activities suspected of constituting threats to the security of Canada, and to report on these to the Government of Canada.

CSIS gathers intelligence information and disseminates it to appropriate government policy-makers. They define a five-step cycle: Government Direction, Planning, Collection, Analysis and Dissemination.

https://www.csis-scrs.gc.ca/bts/ccl-en.php

Introduction - OSINT



Cross Industry Standard Process for Data Mining (CRISP-DM)

Defines a data mining process model that describes commonly used approaches that data mining experts use to tackle problems.

CRISP-DM was conceived in 1996 and was led by five companies: SPSS, Teradata, Daimler AG, NCR Corporation and OHRA. It defines a cycle with six interconnected steps: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment.

Shearer C., *The CRISP-DM model: the new blueprint for data mining*, J Data Warehousing (2000); 5:13—22

European Commission

Autoritas Social Business Intelligence Cycle

European Commission

What? <- Conceptualising brand

What are people talking about?

Introduction
- OSINT

European Commission

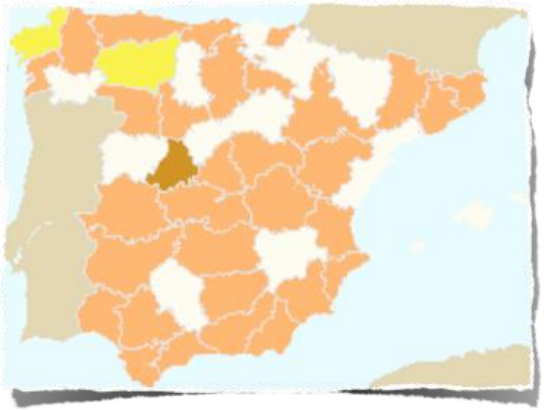
When? <- Crisis management

When is something happening?

Eurostat

European Commission

Where? <- In 2 dimensions



<--- Where does a conversation occur?

Where is the focus of a conversation? --->

Introduction
- OSINT

European Commission

How? <- Not only sentiment analysis

Polarity is only one dimension, emotional, values, SWOT... all of them may answer the HOW question

Eurostat

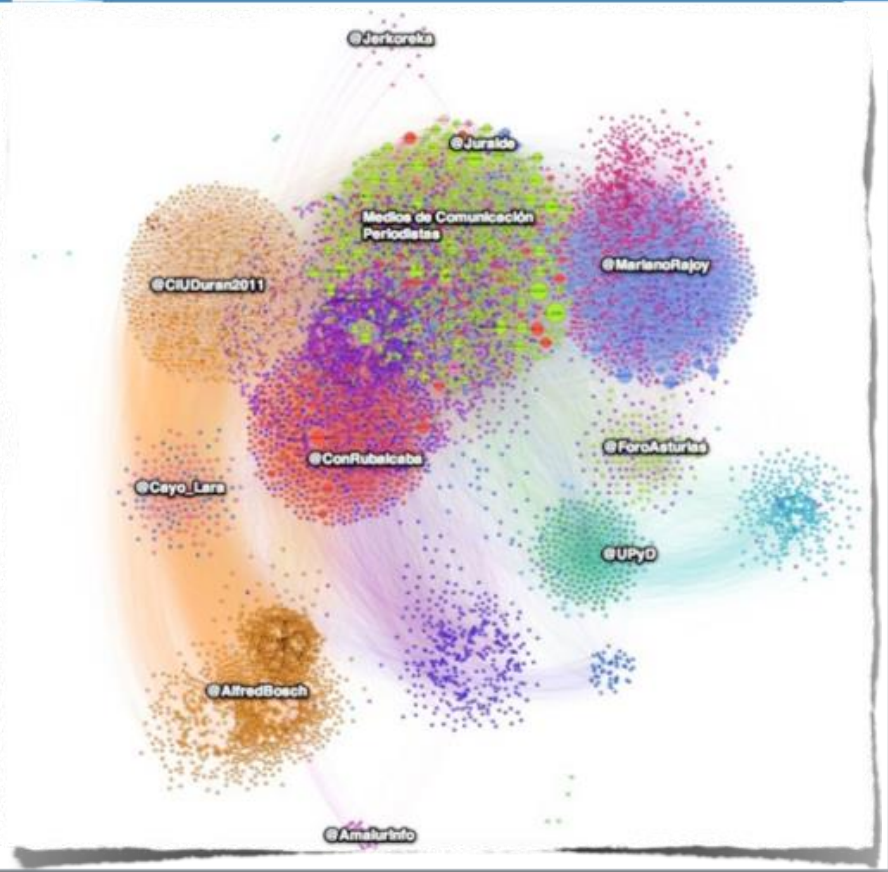
Introduction
- OSINT

European Commission

Who? <- Social Network Analysis

If I want to transmit a message with success, who can help me?

If there is a conflict, who do I have to watch?

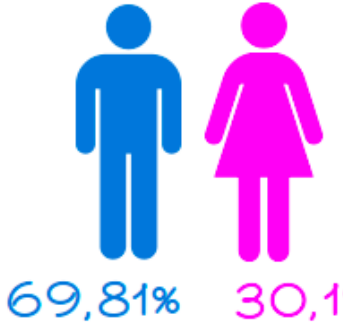


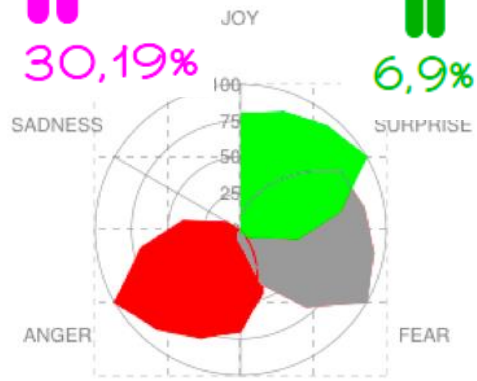
Eurostat

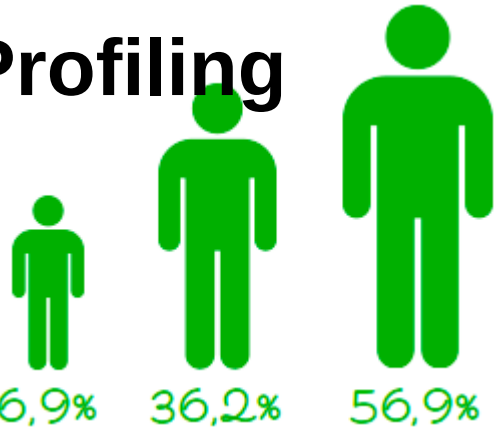
Introduction
- OSINT

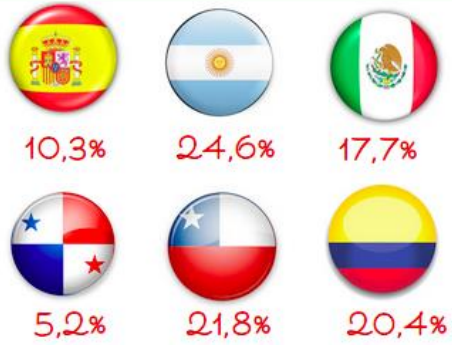
European Commission

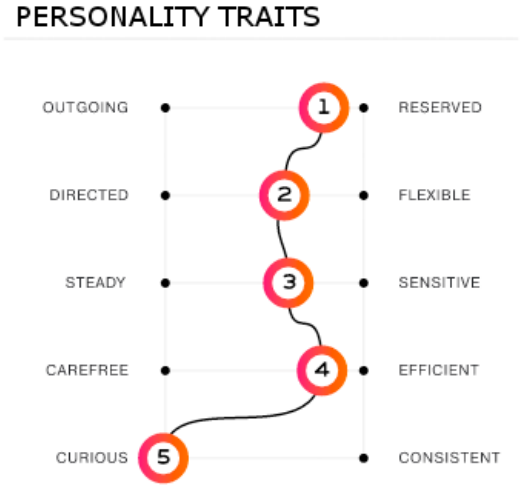
Why? <- Author Profiling











Eurostat

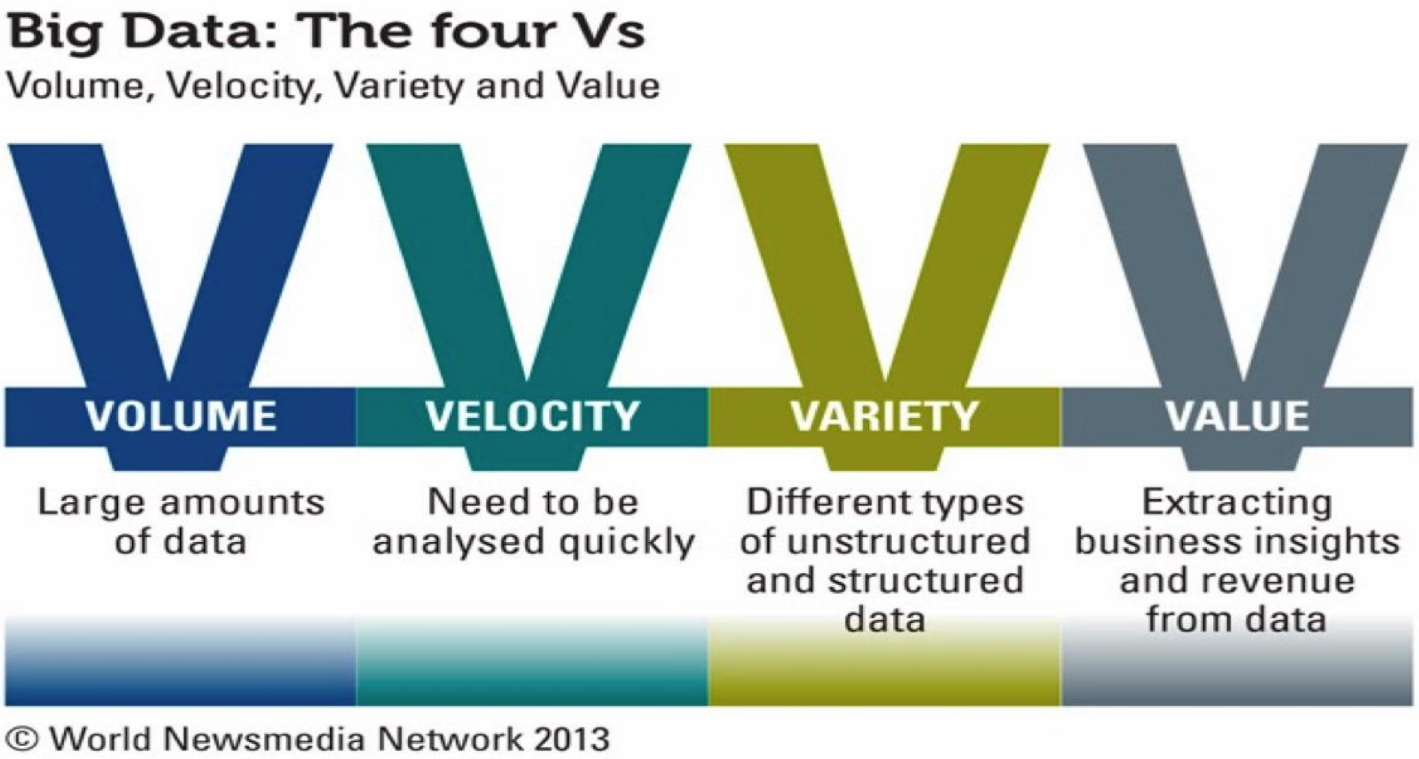
Common tasks

- Understand the business
- Collect, prepare and clean data
- Analyse, evaluate and interpret information
- Disseminate intelligence

Introduction
- OSINT

European Commission

Introduction
- OSINT
- HUMINT
- GEOINT
- MASINT
- SIGINT
- TECHINT
- MEDINT
- CYBINT
- DNINT
- FININT



Introduction
- Infrastructure

European Commission

Microsoft Azure

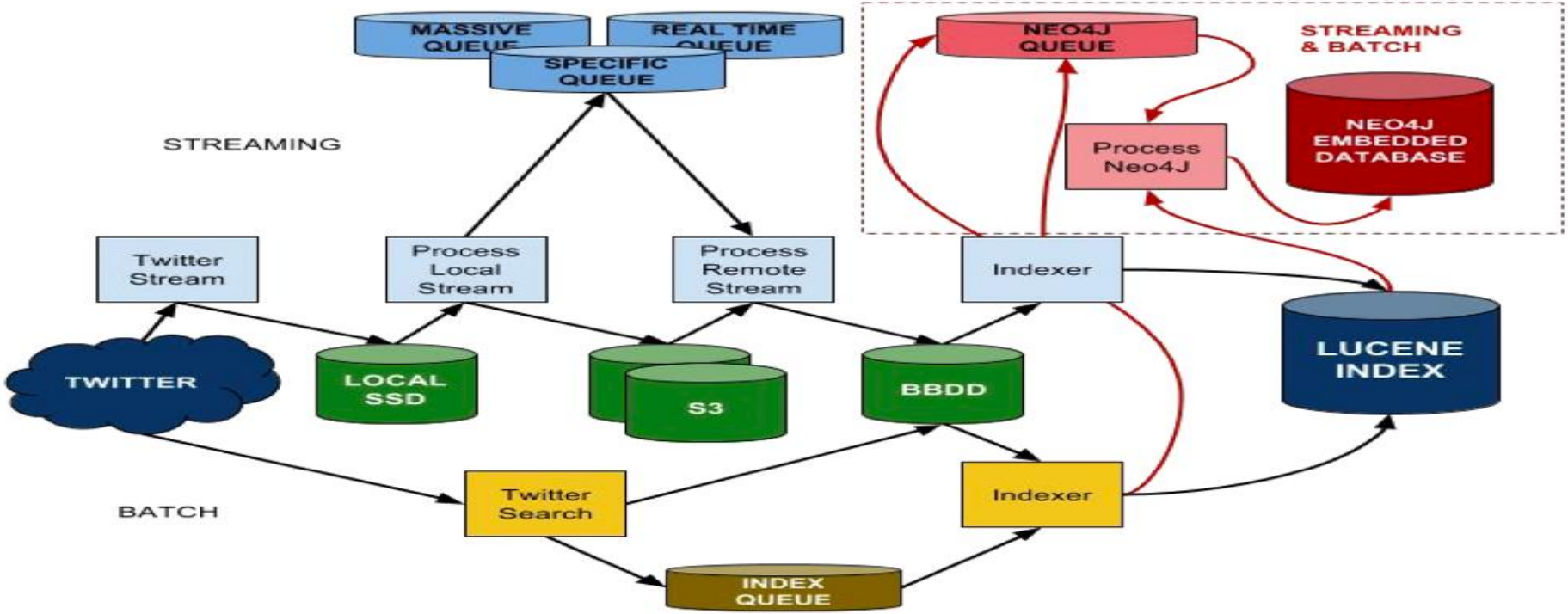
amazon web services™

rackspace
the open cloud company

Google App Engine

Gartner: Magic Quadrant for Cloud Infrastructure as a Service, Worldwide report, Lydia Leong et al. May 18th, 2015

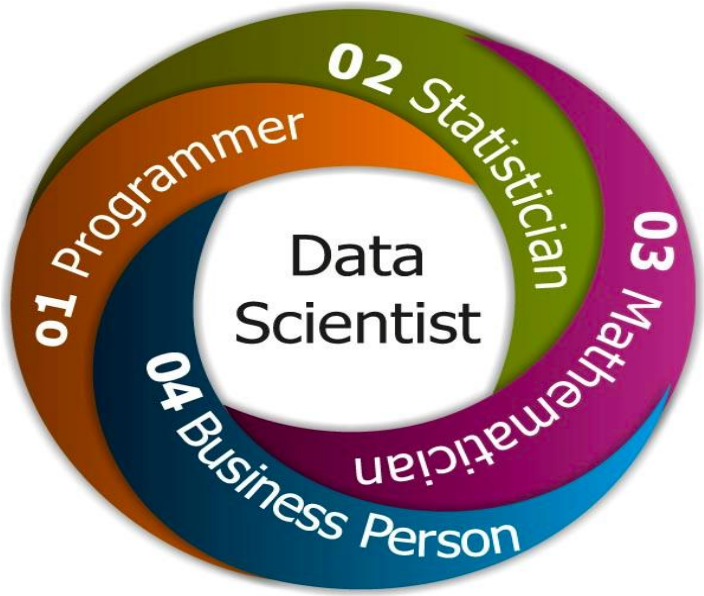
Introduction
- Infrastructure



Introduction
- Technical Skills

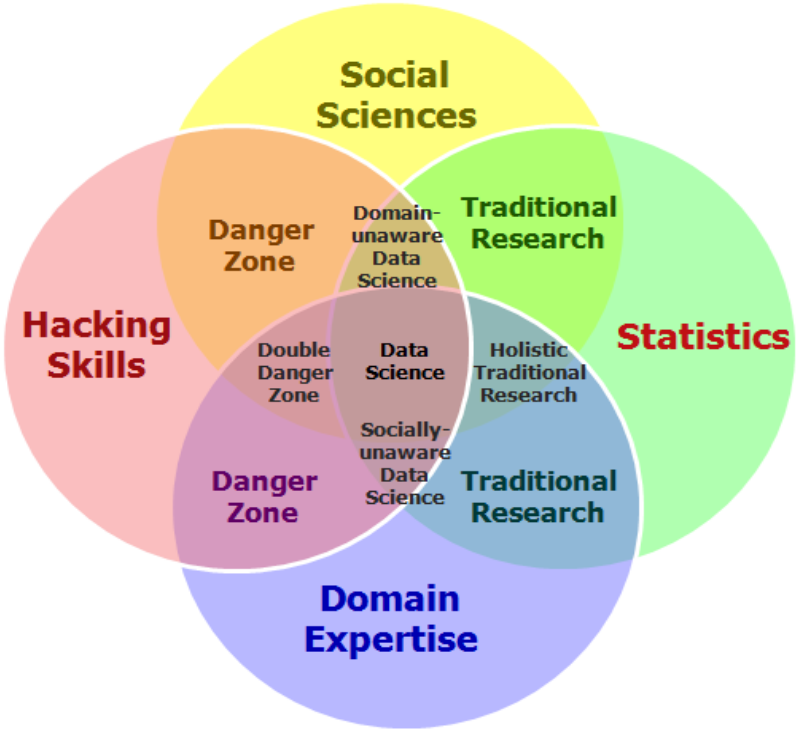
+

Introduction
- Technical Skills



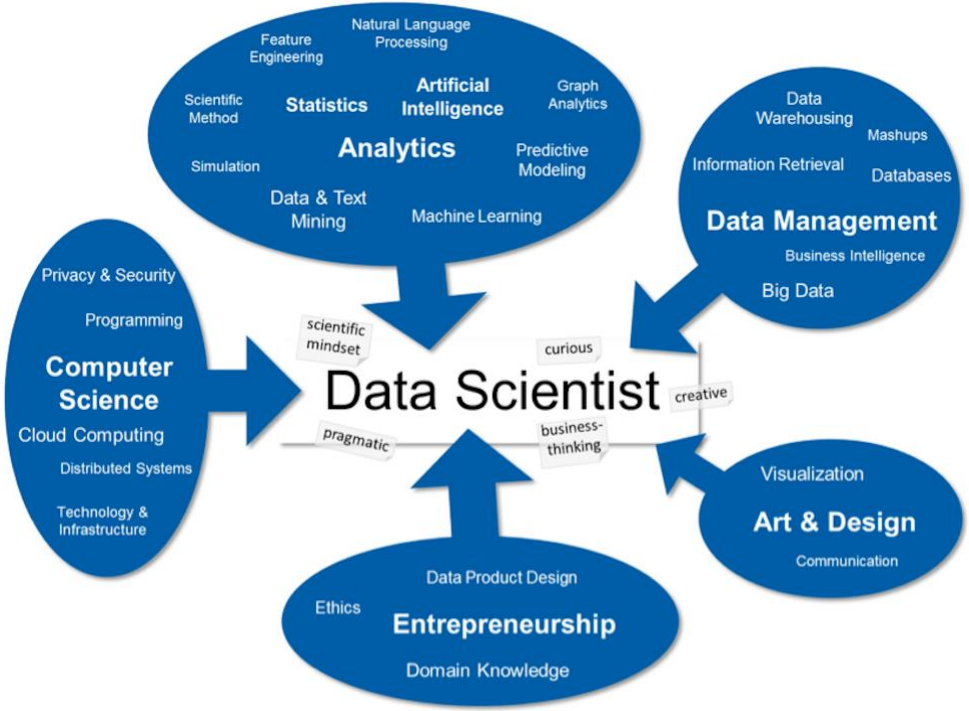
http://www.edureka.co/blog/who-is-a-data-scientist/

Introduction
- Technical Skills



http://www.journaldunet.com/solutions/expert/59530/comment-le-big-data-et-les-data-scientists-valorisent-l-open-source.shtml

Introduction
- Technical Skills



https://blog.zhaw.ch/datascience/the-data-science-skill-set/

Introduction
- Technical Skills

In summary, a good data scientist should be a person able to:

1. understand business fundamentals to translate business needs into data problems;
2. manage computer tools to prepare, retrieve and clean data;
3. master statistics and mathematics to analyse and evaluate information;
4. collaborate with other team members; and
5. communicate results in a proper manner.

Outline

1. Introduction to Network Analysis and Web Analytics
    a. OSINT: Open Source INTelligence
    b. Infrastructure Review
    c. Technical Skills
2. **Information Sources, Retrieval and Extraction**
3. Network Analysis
    a. Graph Construction
    b. Connectivity Degree, Connected Components and Giant Component
    c. Centrality Measures: Influence
    d. Communities
4. Practical examples using R and Gephi

Information sources
- Source types

There are almost infinite different sources, but they can be grouped into:

- Search engines
- RSS channels
- Alerts
- Open data
- Social media

Information sources
- Search engines

Google

bing™

YAHOO!®

Information sources
- **Search engines types**

There are different kinds of search engines:

- General
  - Google, Yahoo!, Bing...
- Thematic
  - Carrot2: http://search.carrot2.org
- Patents
  - National Agencies (e.g. http://consultas2.oepm.es/InvenesWeb)
  - Google Patents (https://patents.google.com)
- Legal
  - The Public Library of Law (http://www.plol.org)
  - National Agencies (e.g. http://www.poderjudicial.es/search/indexAN.jsp)
- ...

Information sources
- RSS channels

European Commission


```
<?xml version="1.0" encoding="UTF-8" ?>
<rss version="2.0">
<channel>
 <title>Título del RSS</title>
 <description>Descripción del RSS</description>
 <link>http://www.sitiodelquesedeseapublicar.com/main.html</link>
 <lastBuildDate>Mon, 06 Jan 2013 00:01:00 +0000 </lastBuildDate>
 <pubDate>Mon, 06 Jan 2013 16:20:00 +0000 </pubDate>
 <ttl>1800</ttl>

 <item>
  <title>Entrada dentro del RSS</title>
  <description>Descripción de la entrada</description>
  <link>http://www.sitiodelquesedeseapublicar.com/enero-2013.html</link>
  <guid>clave única</guid>
  <pubDate>Mon, 06 Jan 2013 17:20:00 +0000 </pubDate>
 </item>

</channel>
</rss>
```

Rich Site Summary or Really Simple Sindication, publish the latest news of the site with full or summarized text and metadata, like publishing data or author's name.

Information sources
- Google Alerts

PRACTICE:

- Creation of alerts with Google Alerts
- Generation of RSS feeds from alerts

Information sources

- Open data

BOLETÍN OFICIAL DEL ESTADO

Information sources
- **Open data**

- National agencies (e.g. http://www.ine.es)
- Eurostats (http://epp.eurostat.ec.europa.eu/portal/page/portal/statistics/search_database)
- U.S. Census Bureau (http://www.census.gov/population/international/data/idb/informationGateway.php)
- U.S. Census Bureau - List to national statistics agencies (http://www.census.gov/population/international/links/stat_int.html)
- World Bank (http://data.worldbank.org)
- United Nations (http://data.un.org)
- CEPAL statistics (http://websie.eclac.cl/infest/ajax/cepalstat.asp)
- Asian-Pacific statistics (http://www.unescap.org/stat/data/swweb_syb2011/DataExplorer.aspx)
- OMPI Patents (http://www.wipo.int/patentscope/search/en/search.jsf)
- OMPI Brands (http://www.wipo.int/madrid/en/romarin)

Information sources
- Social media

Instagram

LinkedIn

Pinterest

WIKIPEDIA
La enciclopedia libre

slideshare

foursquare

Information retrieval
- Practice

- Create a Twitter account
- Create a Twitter app and obtain API credentials
- Retrieve tweets with a R application

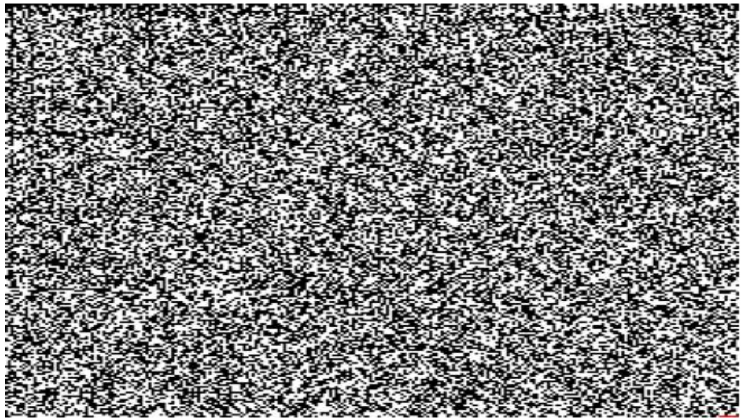
Information extraction
- Practice

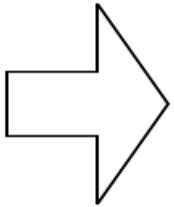
- Extract user mentions with a R application
- Extract hashtags with a R application

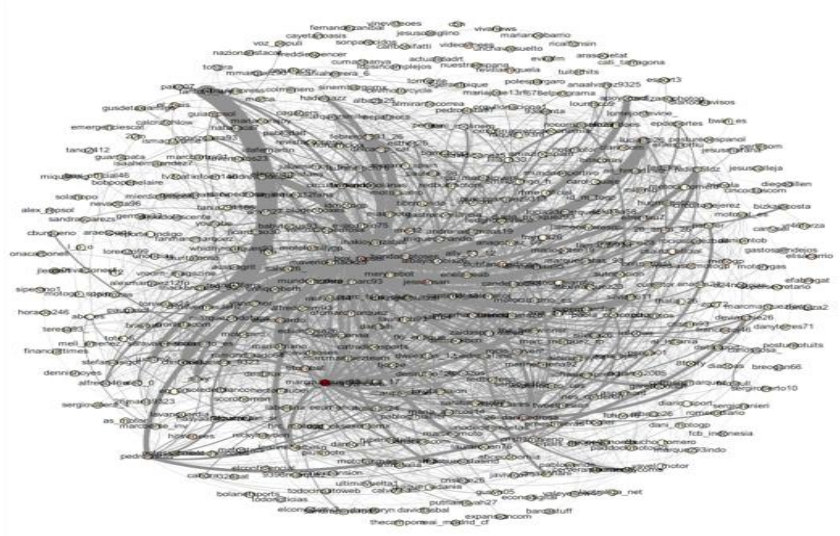
Outline

1. Introduction to Network Analysis and Web Analytics
    a. OSINT: Open Source INTelligence
    b. Infrastructure Review
    c. Technical Skills
2. Information Sources, Retrieval and Extraction
3. **Network Analysis**
    a. Graph Construction
    b. Connectivity Degree, Connected Components and Giant Component
    c. Centrality Measures: Influence
    d. Communities
4. Practical examples using R and Gephi

Network Analysis







Network Analysis

Name	Age	Gender	Profession
Fran	38	Male	Computer Scientist
Mary	28	Female	Lawyer
Michael	67	Male	Retired
Helen	17	Female	Student

Demographic data of users from a social network

Network Analysis

	Fran	Mary	Michael	Helen
Fran	-	1	0	1
Mary	0	-	1	0
Michael	1	0	-	1
Helen	1	1	0	-

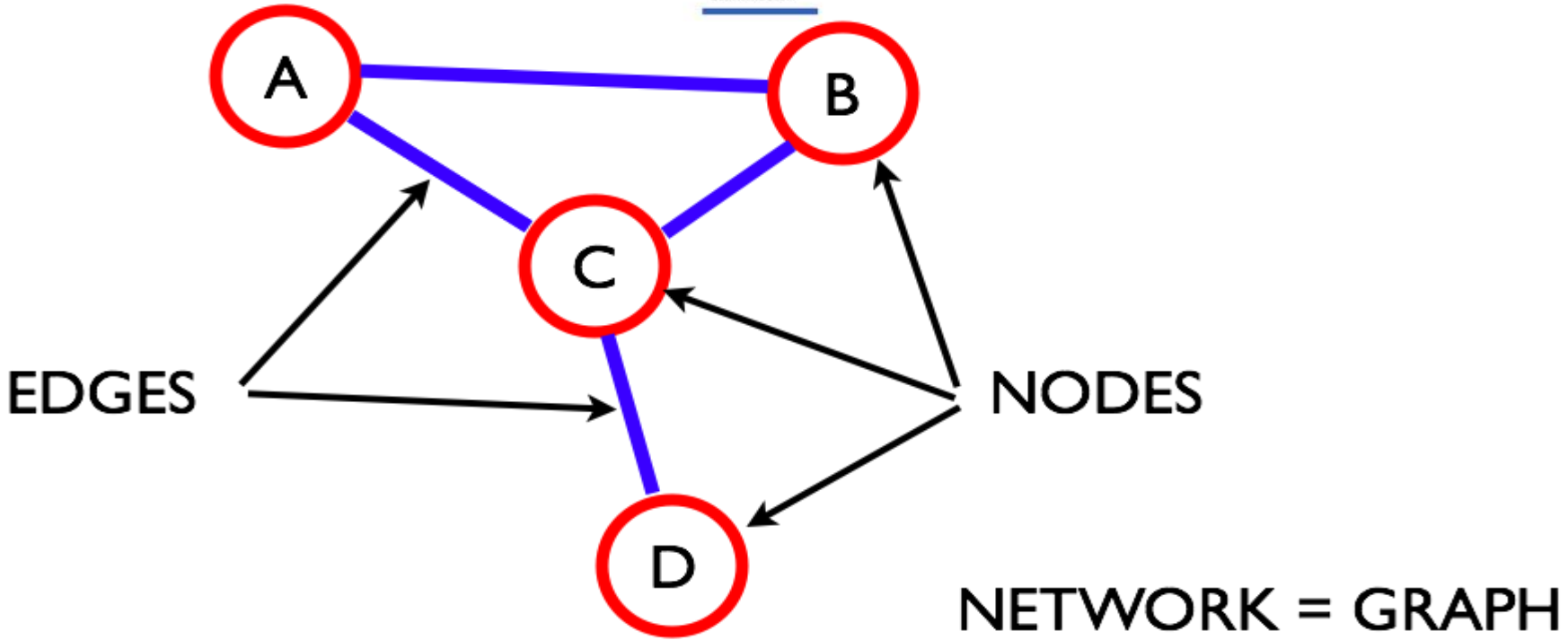
Who follows whom in a social network?

Network Analysis
- Representations

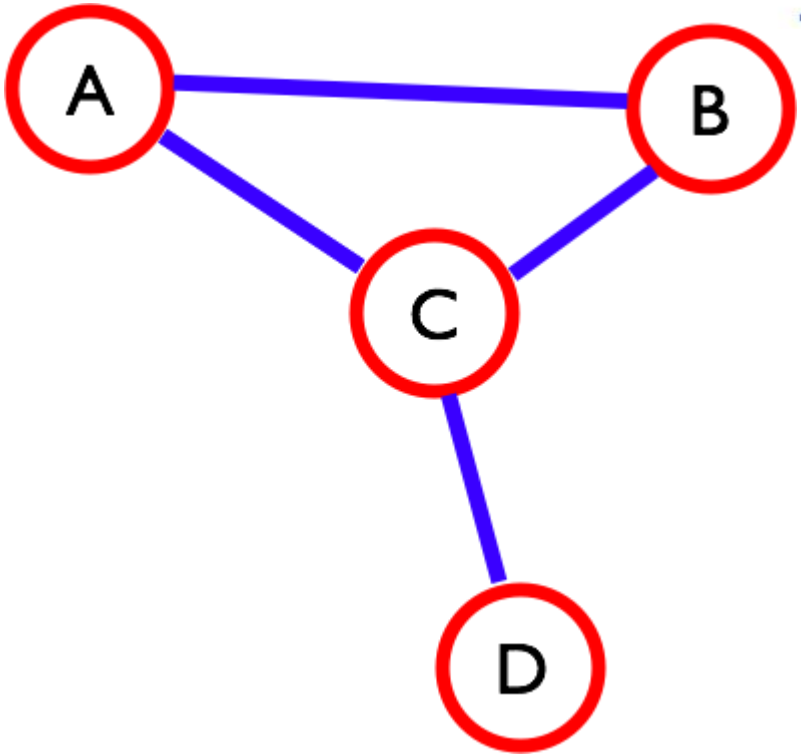
There are different kinds of network representations:

- Graph
- Adjacency matrix
- Linked list
- Adjacency list

Network Analysis
- Graph

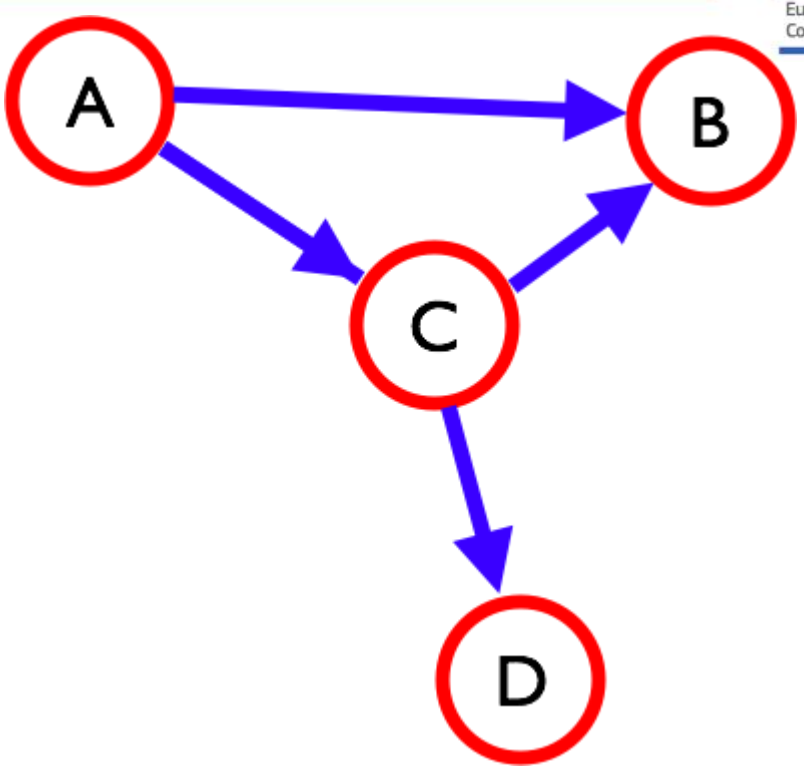


Network Analysis
- **Undirected graph**



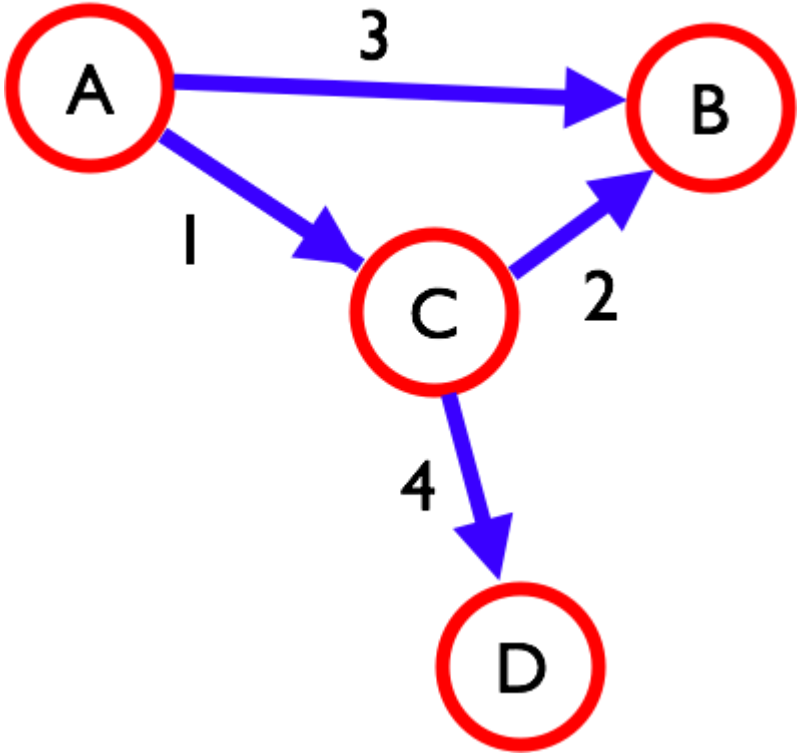
Examples:
- Facebook friends
- Classmates
- Relatives

Network Analysis
- Directed graph



Examples:
- Twitter followers/friends
- Blogosphere
- Family tree

Network Analysis
- Weighted graph

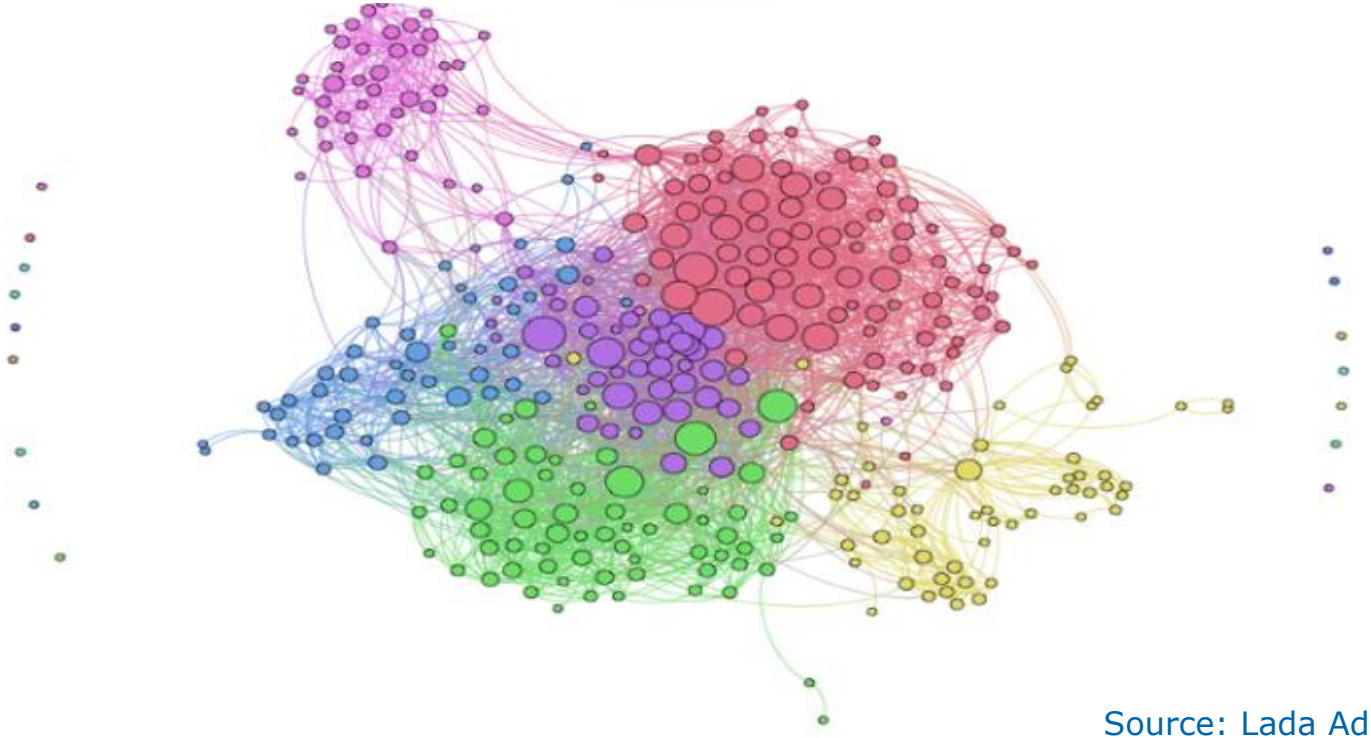


Examples:
- Network iterations
- Distance among nodes
- Affinity degree
- Traffic modeling

Network Analysis - Nomenclature

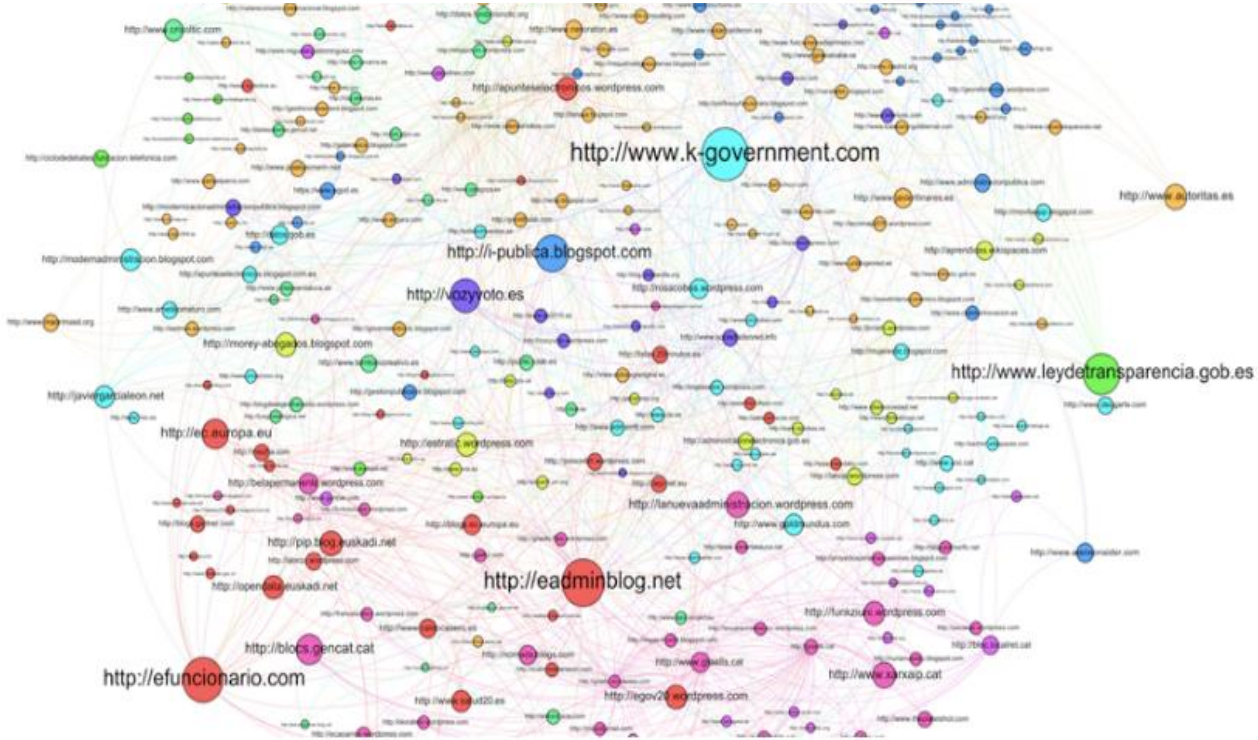
Discipline	Points	Lines
Mathematics	Vertex	Axes, Arcs
Informatics	Nodes	Links
Physics	Sites	Connections
Sociology	Actors	Ties, Relations

Network Analysis
- Facebook friends



Source: Lada Adamic - School of Information, University of Michigan

Network Analysis
- Blogosphere



Source: Carlos Guadian
http://www.k-government.com

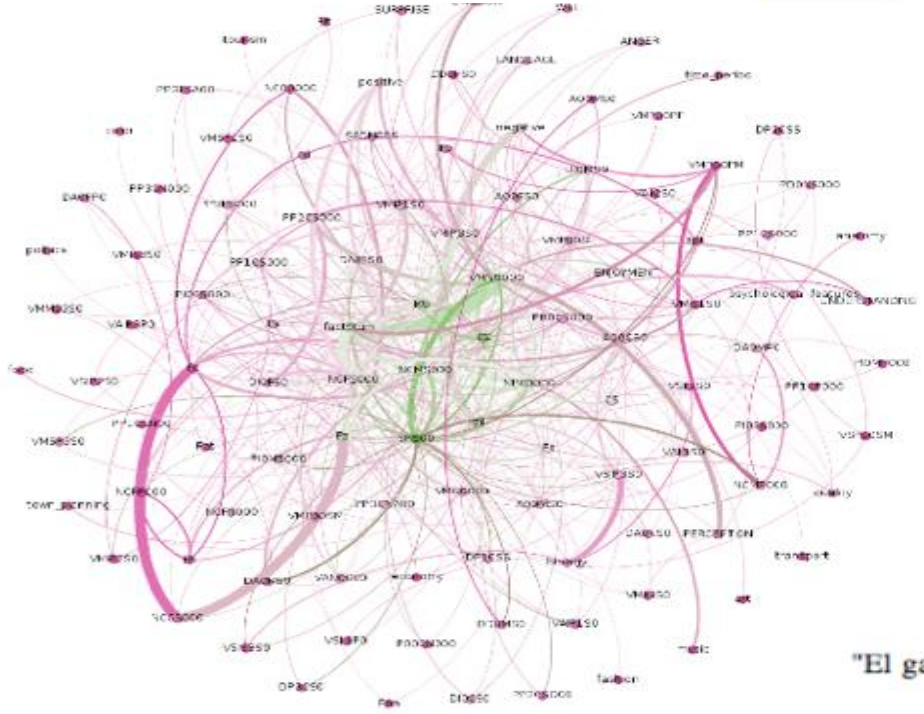
Network Analysis
- Twitter network

European Commission

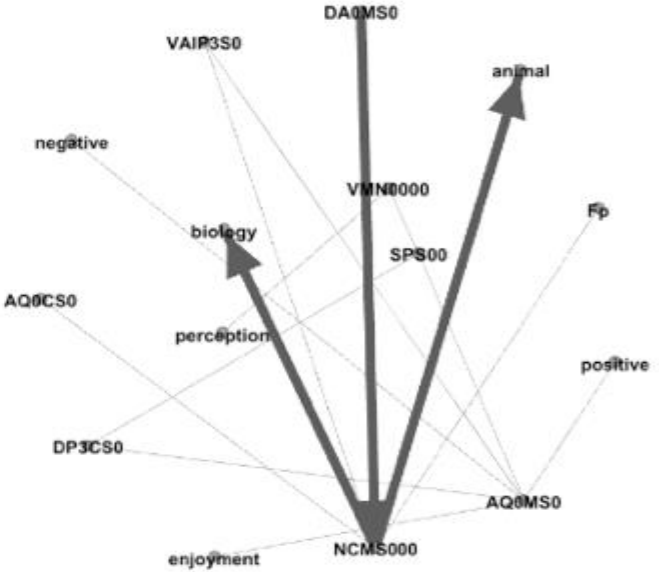
Structural: Followers/Friends
Behavioural: Mentions/RTs

Source: Francisco M. Rangel

Network Analysis
- Discourse analysis

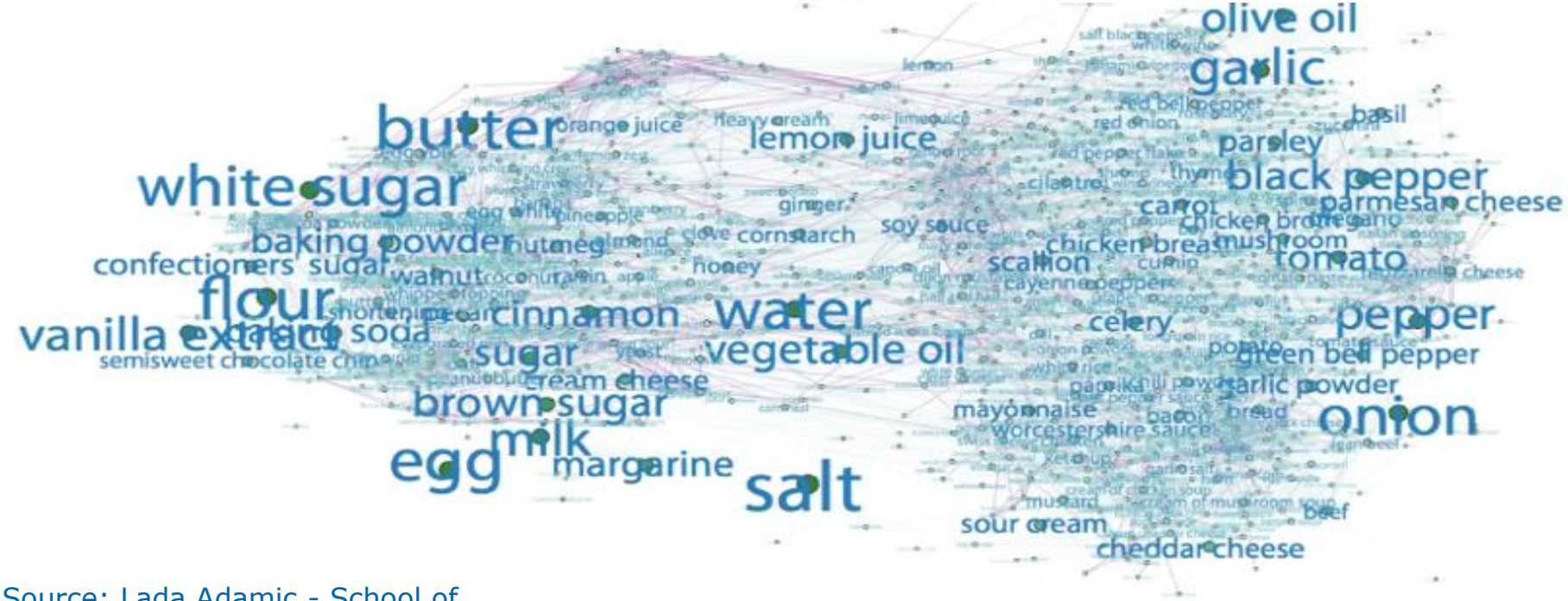


Source: Francisco M. Rangel
© Autoritas Consulting



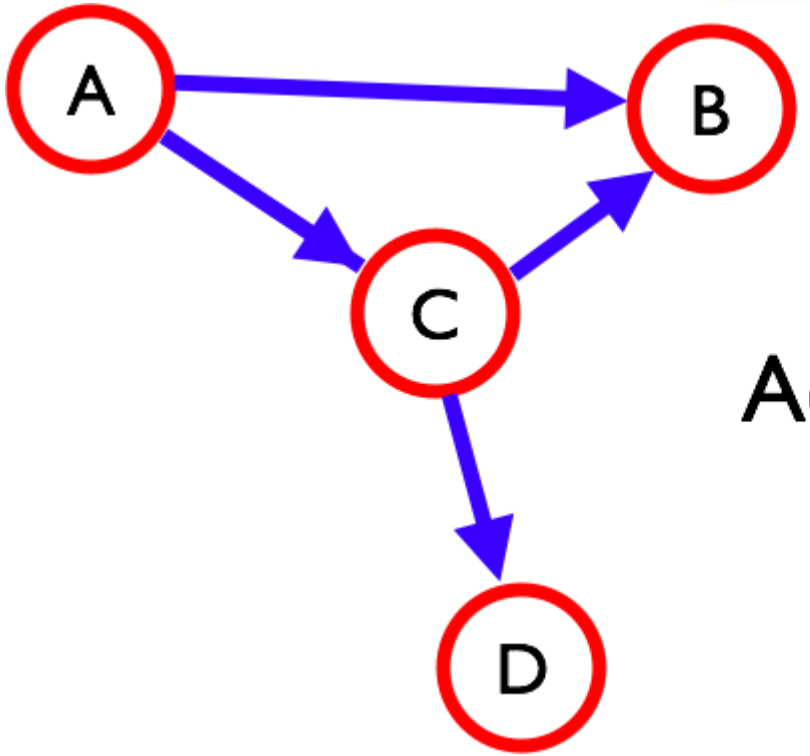
"El gato está contento de ver a su increíble amigo el perro."

Network Analysis
- Food ingredients



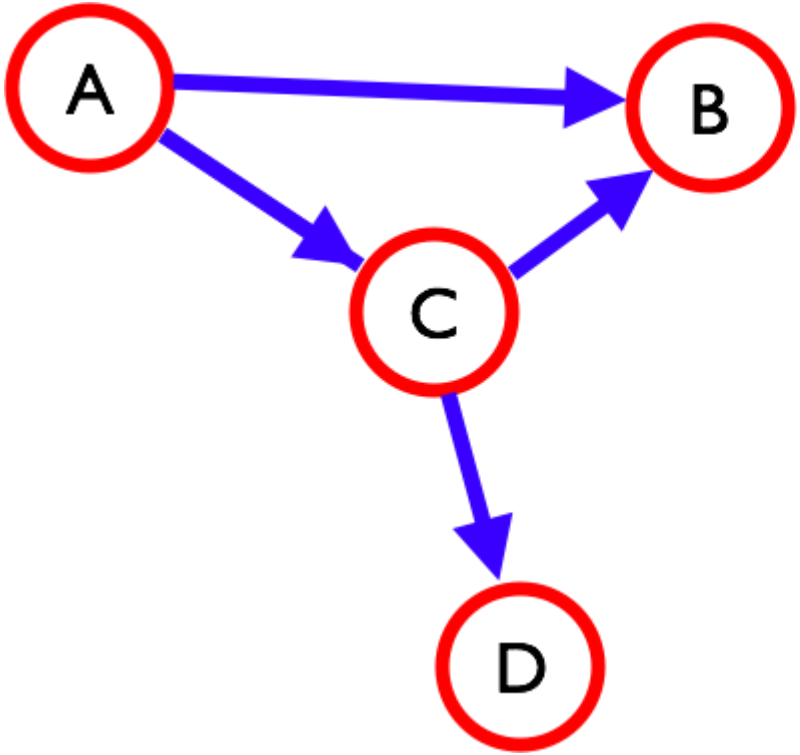
Source: Lada Adamic - School of Information, University of Michigan

Network Analysis
- Adjacency matrix



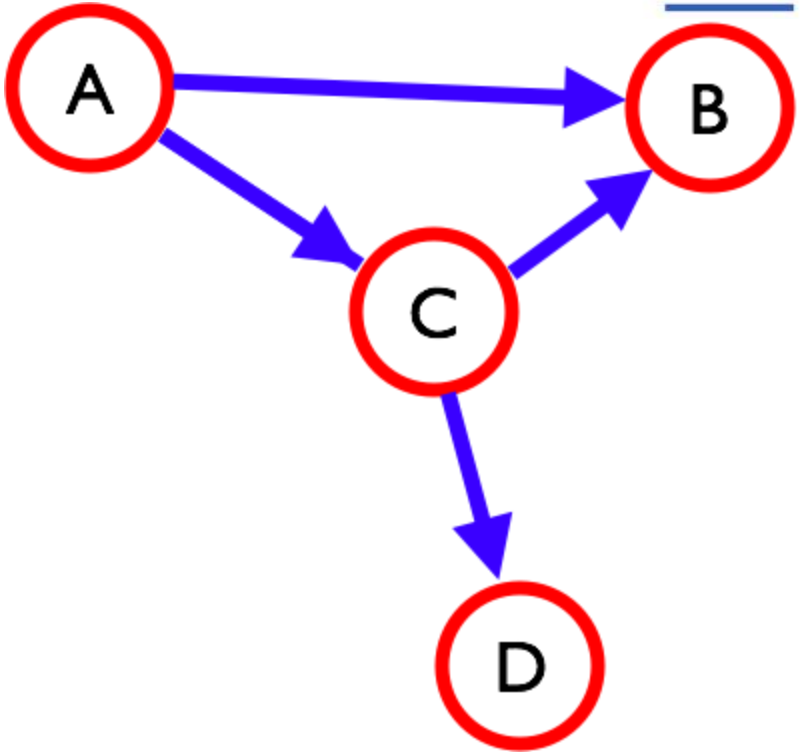
$$\text{Adj.} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

Network Analysis
- Linked list



- A, B
- A, C
- C, B
- C, D

Network Analysis
- Adjacency list



- A: B, C
- B:
- C: B, D
- D:

Network Analysis - Properties

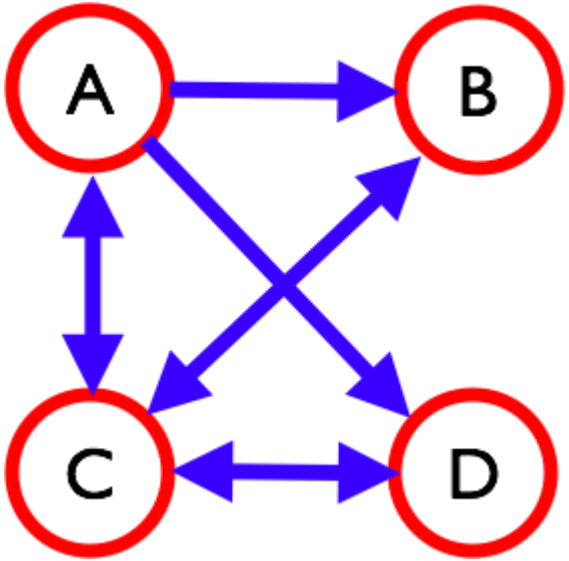
Representation	Advantage
Graph	Visual
Adjacency matrix	Centrality calculations
Linked list	Economic
Adjacency list	Neighborhood calculations

Network Analysis
- Motivation

What are the motivations for using these representations?

- Concrete and formal representations
- Use of computers
- Ability to generate new questions

Network Analysis
- New questions



- Nodes represent people
  - A and C are males
  - B and D are females
- Links represent that somebody likes somebody
- A and C like more people than B and D
- Might there be a pattern? For example, might men like more people than women?

Network analysis - GEXF

European Commission

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf xmlns="http://www.gexf.net/1.2draft" version="1.2">
<meta lastmodifieddate="2013-12-17">
<creator>Cosmos</creator>
<description>Twitter Hashtags Relationship</description>
</meta>
<graph mode="static" defaultedgetype="undirected">
    <nodes>
        <node id="#isla" label="#isla" />
        <node id="#tonicladerataxi" label="#tonicladerataxi" />
        <node id="#noelvolem" label="#noelvolem" />
        ...
    </nodes>
    <edges>
        <edge id="1" source="#palmaboatshow" target="#náutica" weight="6"></edge>
        <edge id="2" source="#sansalvador" target="#mejicanos" weight="1"></edge>
        <edge id="3" source="#ioibiza" target="#allureoftheseas" weight="1"></edge>
        ...
    </edges>
</graph>
</gexf>
```

Metadata information such as the creator and the description

static vs. dynamic. directed vs. undirected

Nodes definition section. Nodes are defined as a list of elements.

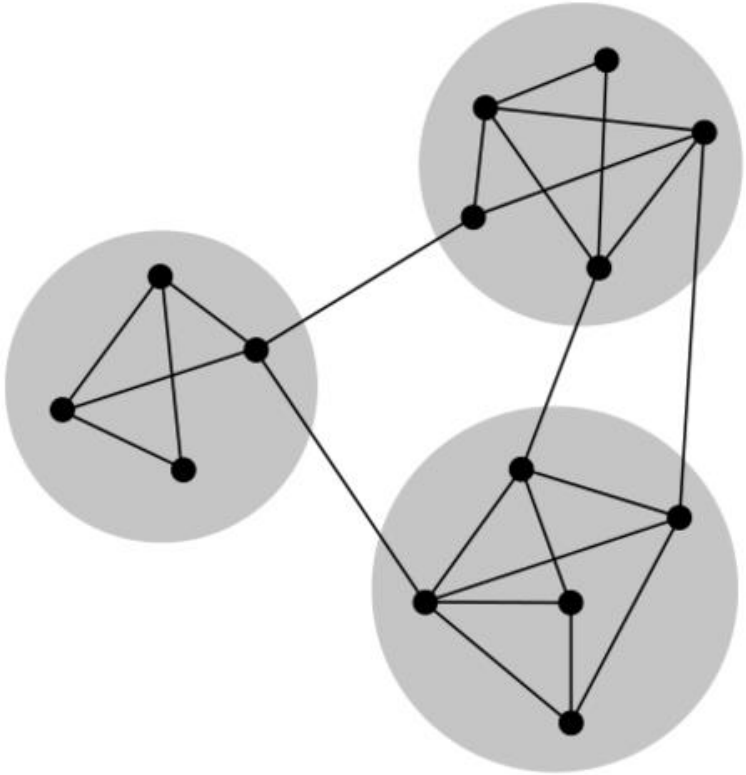
Edges definition section. Edges are defined as pairs of source and destination nodes, with a possible weight.

all the xml elements should be closed

Network analysis
- Practice

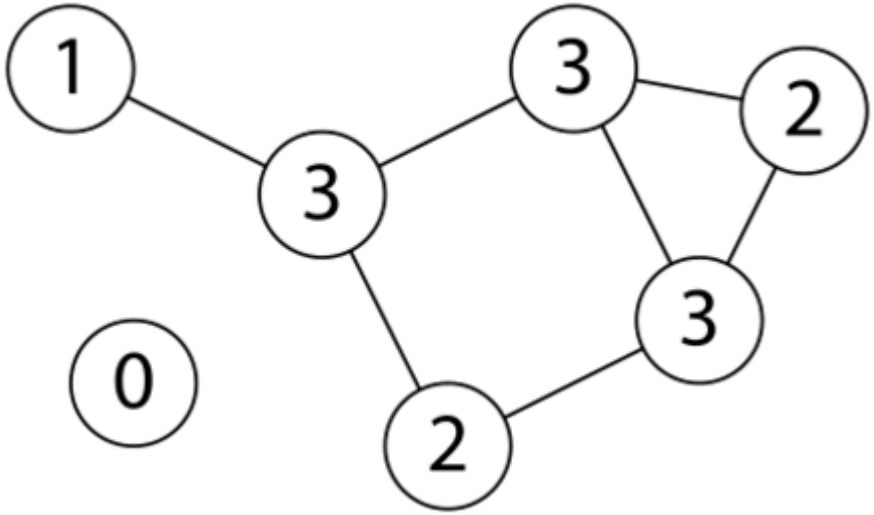
- Twitter conversation GEXF with R
- Hashtag-User GEXF with R

Network Analysis
- Connectivity Degree



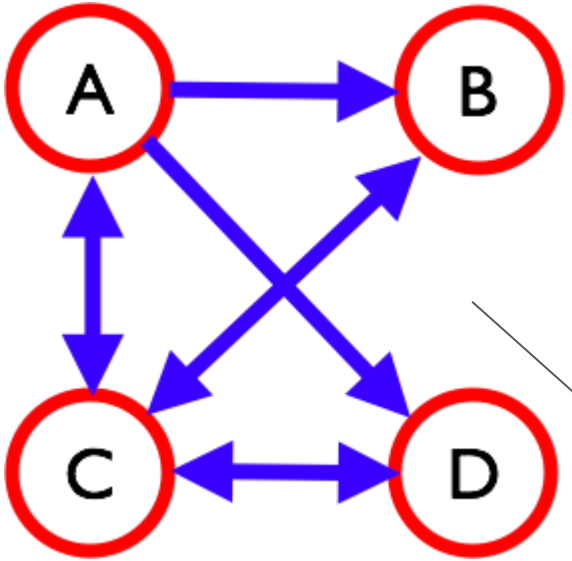
The minimum number of nodes or edges that need to be removed to disconnect the remaining nodes from each other.

Network Analysis
- Connectivity Degree



Each node degree is the number of edges between this node and other nodes in the graph.

Network Analysis
- Connectivity Degree

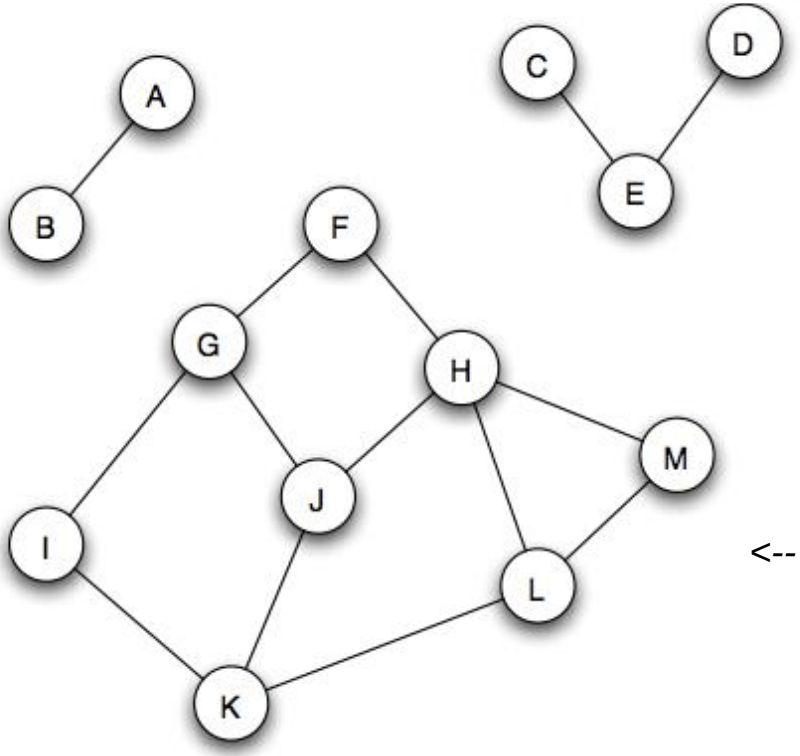


In directed graphs, there are two degrees:

- Indegree or number of edges that arrives to the node.
- Outdegree or the number of edges that leave the node.

What's the indegree and outdegree of each node?

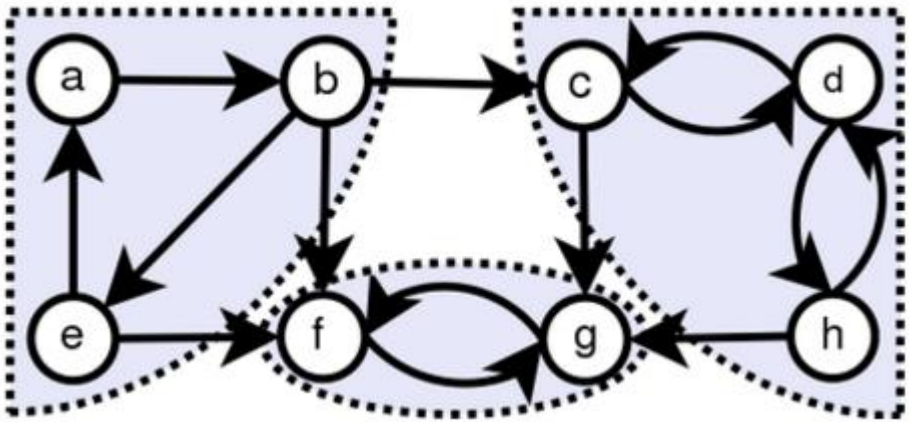
Network Analysis
- Connected Components



A connected component is a subgraph in which any two vertices are connected to each other by paths.

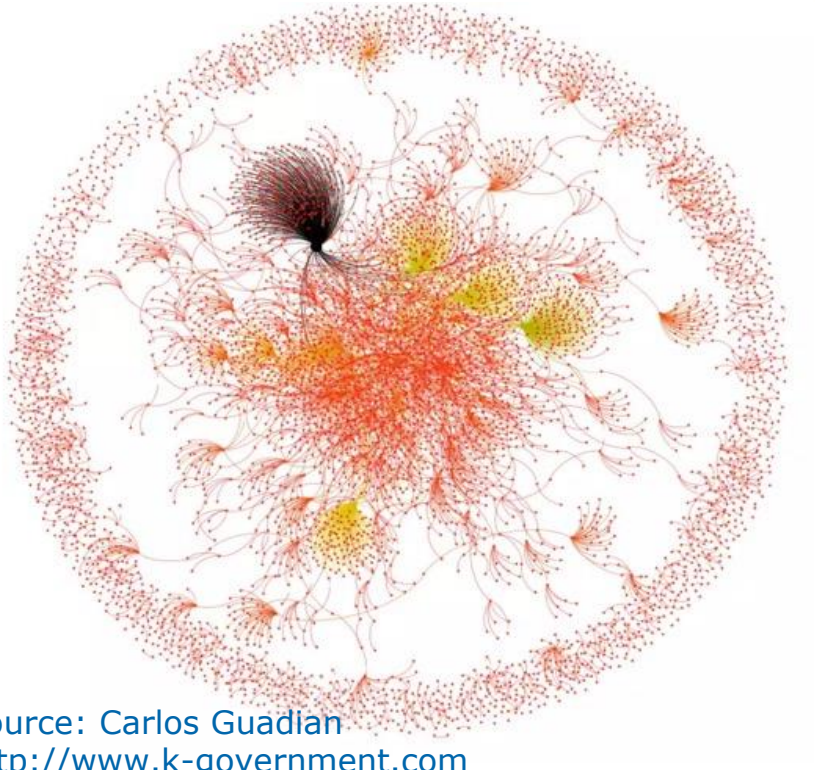
<-- *How many connected components does this graph have?*

Network Analysis
- Connected Components



<-- *And in this one? Why?*

Network Analysis
- Giant Component



Source: Carlos Guadian
http://www.k-government.com

A giant component is a connected component that contains a constant fraction of the entire graph's vertices.

Network Analysis
- Centrality Measures

Centrality measures indicate the most important vertices within a graph:

- The most influential person in a social network.
- The most critical nodes in a infrastructure (such as Internet).
- The highest spreaders of disease.

The word “important” has many different meanings, such as different centrality measures:

- Degree centrality
- Closeness centrality
- Betweenness centrality
- Eigenvector centrality
- ...

Network Analysis
- Centrality: Eigenvector

Eigenvector centrality measures the influence of a node in the graph depending on the influence of other nodes connected to it. In other words, given a node, high-scoring nodes connected to it contribute more than low-scoring ones. It is a recursive measure.

Given a graph and its adjacency matrix $A=(a_{v,t})$ where $a_{v,t}$ is 1 if a node v is linked to a node t, and 0 otherwise, we can calculate the eigenvector centrality score of the node v as:

$$x_v = \frac{1}{\lambda} \sum_{t \in M(v)} x_t = \frac{1}{\lambda} \sum_{t \in G} a_{v,t} x_t$$

Intuitively, the more influential nodes are connected with v, the more influential v is.

Network Analysis
- Centrality: Betweenness

Betweenness centrality counts the number of times a node is part of the shortest path between each other pair of nodes in the graph.

In other words, the betweenness centrality of a node v is the ratio of all shortests paths from one node to any other node in the graph that pass through v.

Mathematically:

$$C_B(v) = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}}$$

The sum of all the shortest paths from s to t that pass through v divided by all the shortest paths from s to t.

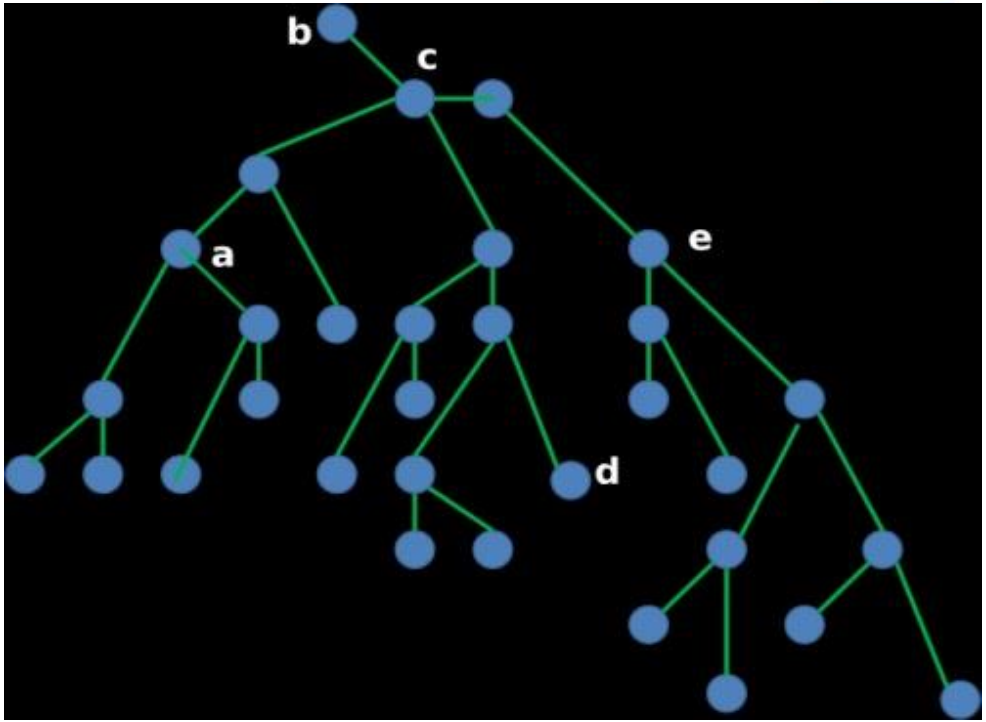
Intuitively, the betweenness measures how important a node is on the basis of how many other nodes depend on it to be connected.

Network Analysis
- Eigenvector vs. betweenness

In the Middle Age, who had more betweenness? and eigenvector?

1. The **peddler**, who knows every mayor and captains of the guard, major retailers and supply officers of all cities, castles, monasteries, baronies and counties through which it passes.
2. The **count**, who overlooks the main castle and all the nobles of the area visit him, and who knows the leading citizens of his county.
3. The **night shift guard**, who must sleep during the day and who is not able to find partner for not meeting people.
4. The **bishop**, who is related to the abbots of all the monasteries and major priorates of his bishopric, that goes beyond any county or local fiefdom.

Network Analysis
- Eigenvector vs. betweenness

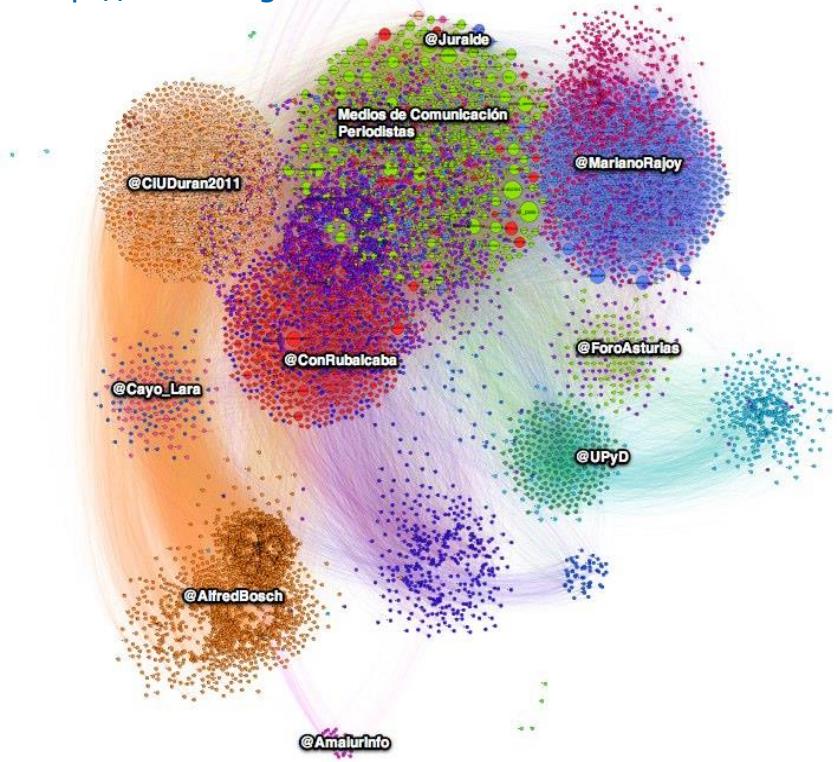


What is the node with the highest betweenness?

http://www.joserodriguez.info/bloc/ayuda-para-entender-grafos-que-es-betweenness-intermediacion/
http://www.joserodriguez.info/bloc/ayuda-para-entender-grafos-ii-que-es-el-eigenvector-relevancia/

Network Analysis - Communities

Source: Carlos Guadian
http://www.k-government.com



A network is said to have a community whether its nodes can be easily grouped into sets of nodes such that each set of nodes is densely connected internally.

Network Analysis
- Practice

- Twitter conversation analysis
- User-hashtags analysis

References

- Social Network Data Analytics. Charu C. Aggarwal. Springer.

  http://www.springer.com/us/book/9781441984616

- Networks, Crowds and Markets: Reasoning about a Highly Connected World. David Easly and Jon Kleinberg. Cambridge University Press.

  https://www.cs.cornell.edu/home/kleinber/networks-book/