

Ioana-Brîndușa Bogdan

Masterand

Faculty of Mathematics and Informatics

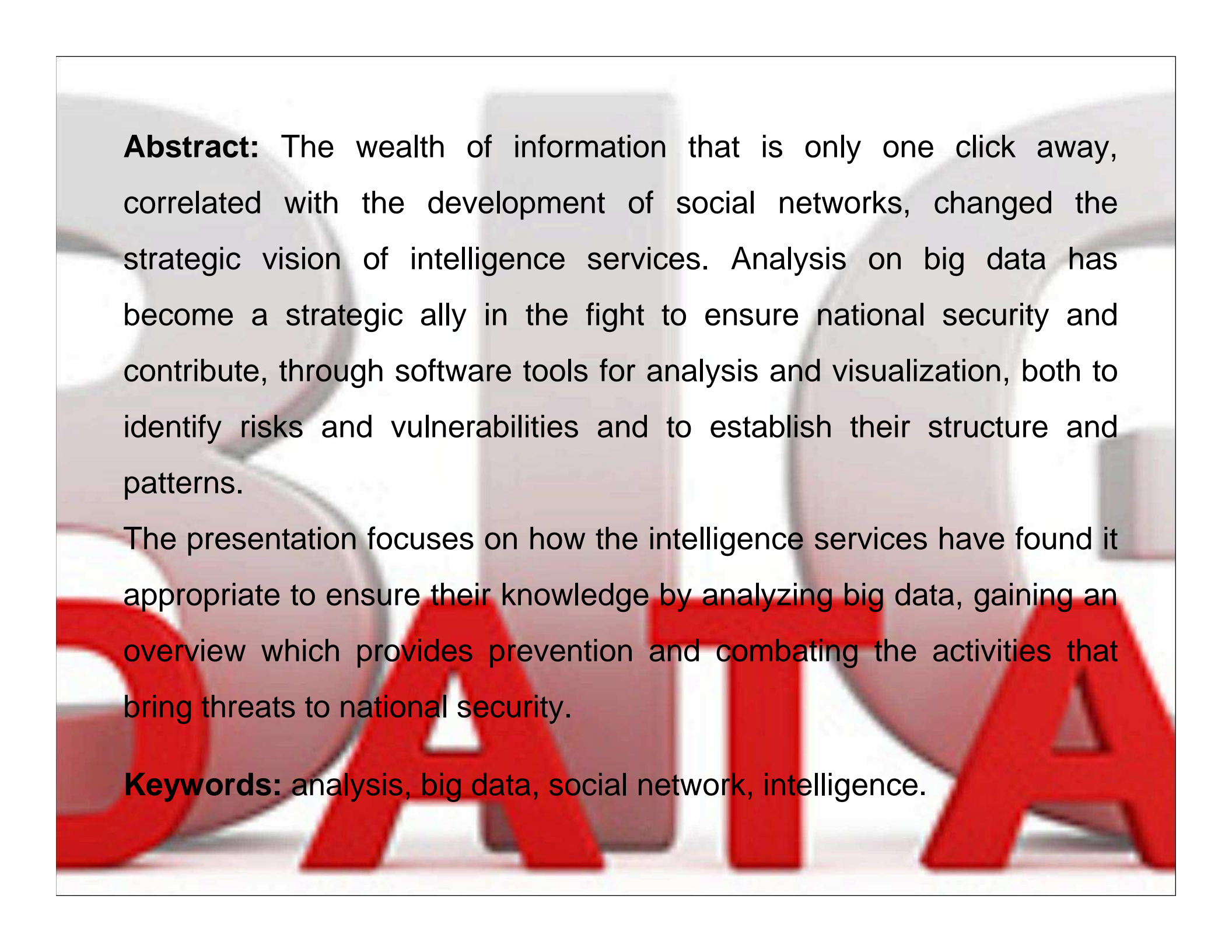
Spiru Haret University

Bucharest, Romania

bicuoana@yahoo.com



THE BENEFITS OF BIG DATA ANALYTICS



Abstract: The wealth of information that is only one click away, correlated with the development of social networks, changed the strategic vision of intelligence services. Analysis on big data has become a strategic ally in the fight to ensure national security and contribute, through software tools for analysis and visualization, both to identify risks and vulnerabilities and to establish their structure and patterns.

The presentation focuses on how the intelligence services have found it appropriate to ensure their knowledge by analyzing big data, gaining an overview which provides prevention and combating the activities that bring threats to national security.

Keywords: analysis, big data, social network, intelligence.

cheap memory



fast Internet connections, and obsessively used

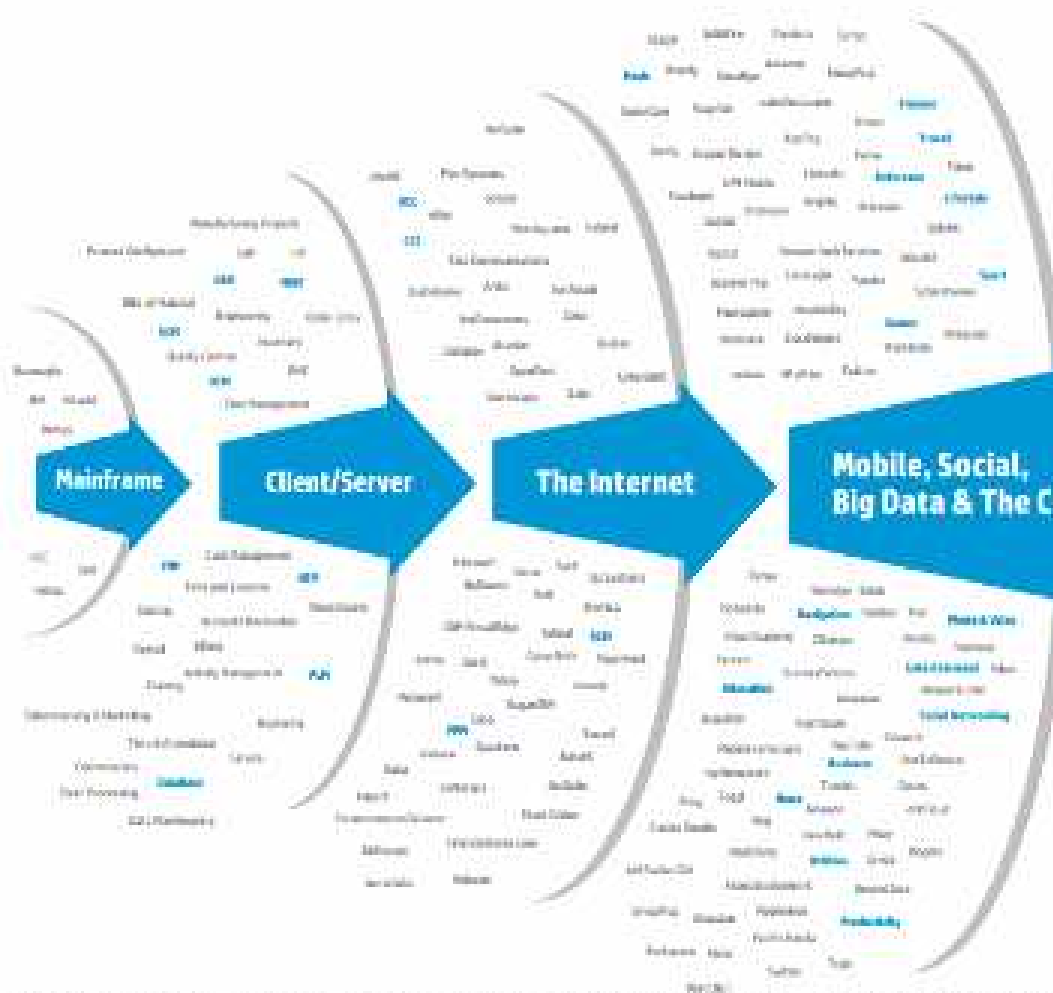


sensor-laden smart-phones



massive datasets

Let's see how much data we are generating every second, of every minute, of every hour, of every day, of every month, of every year.



Every 60 seconds



98,000+ tweets



695,000 status updates



11 million instant messages



698,445 Google searches



168 million+ emails sent



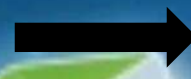
1,820TB of data created



217 new mobile web users



people communicate



reveal significant information

views
sentiments
likes and
dislikes
much more



intelligence services



capture the information



analyze and interpret



valuable information

BIG

a single conversation not reveal much



millions of conversations



identify

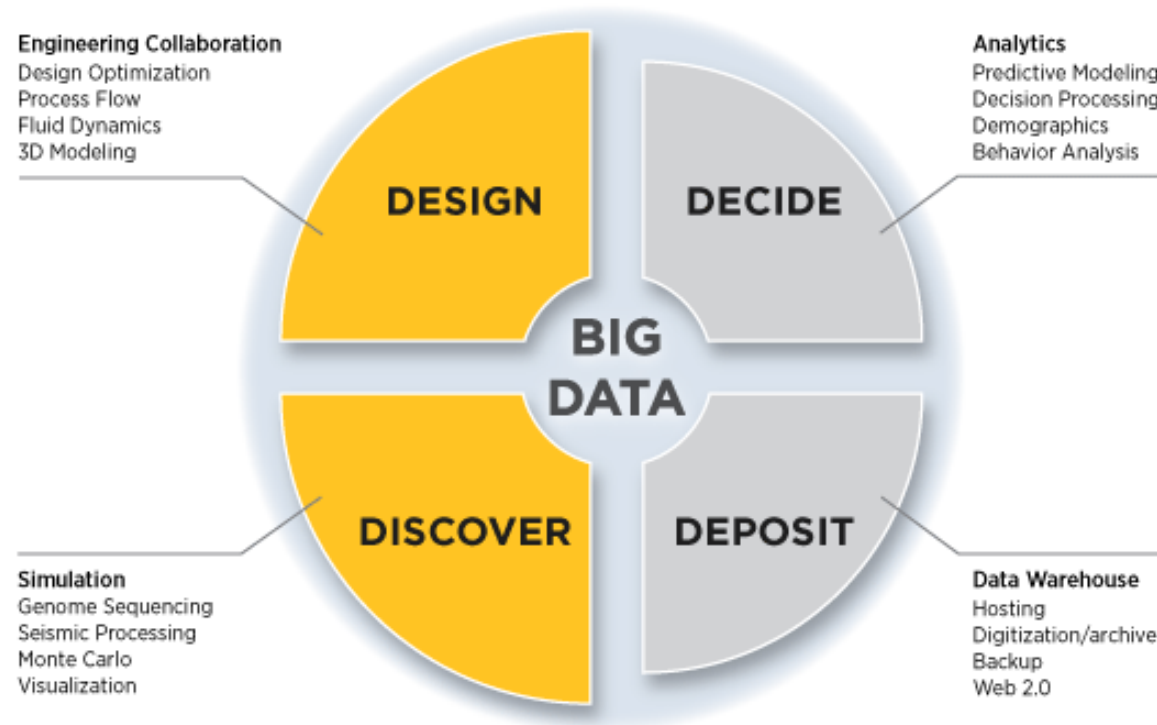


**trends, issues, relationships
target individuals with
relevant**

DATA

No precise definition of "big data" exists, but a good rule of thumb is data sets too large to fit in main memory on a single machine.

Big Data Application Segments



Four distinct applications segments comprise the big data market.

Big Data analytics are the future of providing security.



Intelligence services have to use analysis instruments to unlock the knowledge from these vast numbers of natural language to deliver previously un-heard of levels of big-data analysis.

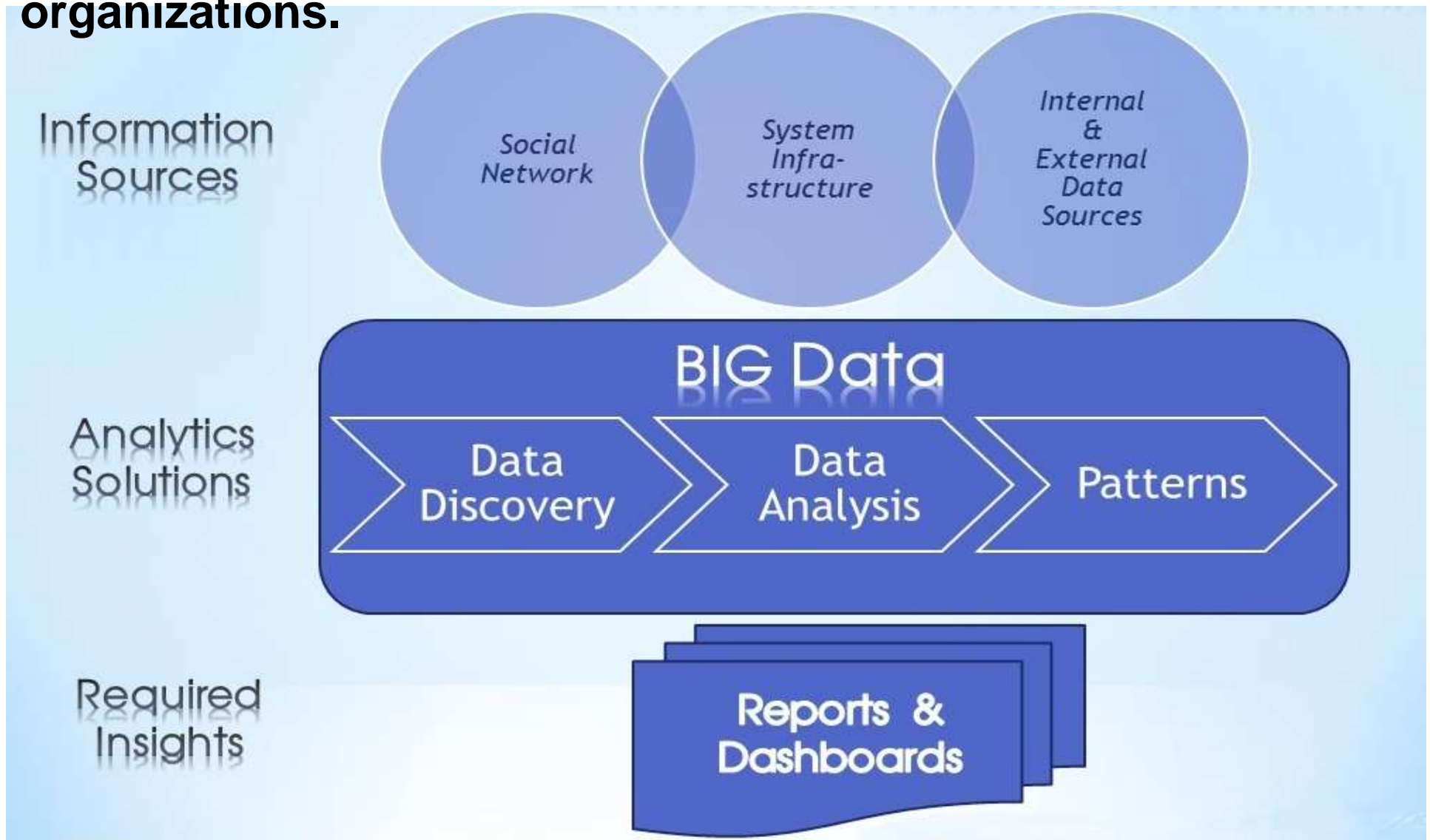


They are getting to know more about how you move around your neighborhood and your city, about your daily habits and schedule.



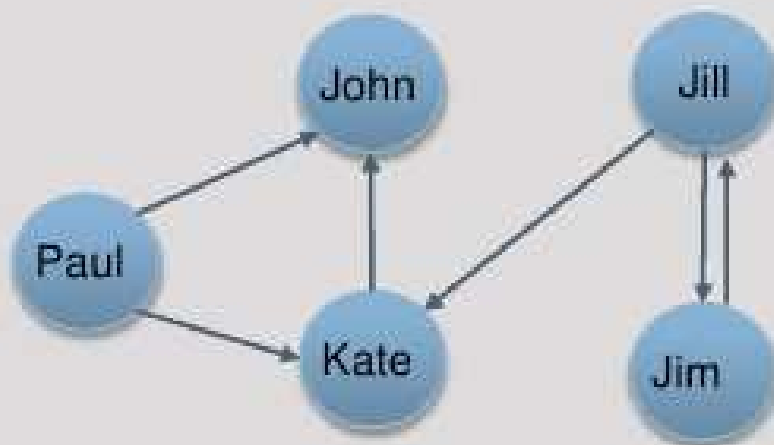
**“where were you at 11:15 last night?”
may be consigned to history.**

Social Network Analysis/ SNA is a mathematical method for “connecting the dots”. SNA allows us to map and measure complex, and sometimes covert, human groups and organizations.

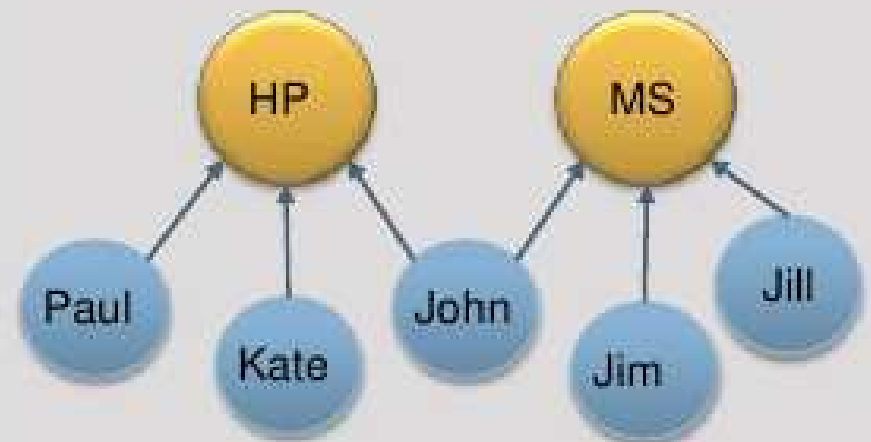


Basic Types of Networks

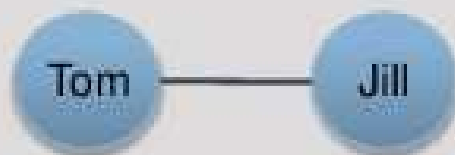
Adjacency
(e.g. friendship nets)



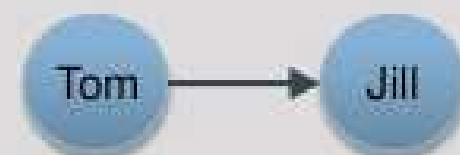
Affiliation
(e.g. employer-employee nets)



Undirected
(e.g. Facebook friends)



Directed
(e.g. Twitter followers)



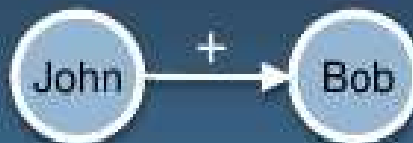
A closer look at links

Binary



- **Link or no link?** (1 or 0)
Typical friendship networks – links exist or they don't, Bob and John are friends or they aren't.

Signed



- **Positive or negative** (+, - or 0)
You can get signed links by asking John "Do you like Bob, dislike him or are neutral about him?"

Valued



- **Weighted links** (each link is assigned a number)
Weights can represent the strength/duration of a relationship. "How many times a week do you call Bob?"

Multiplex



- **Multiplex** (more than one type of link)
Add different relationships between the same set of nodes.

Distance in Networks

Walk



- A route through the network from one node to another. No restraints on repeated nodes or links: A-B-C-D-A-B-C-E is a walk.

Trail



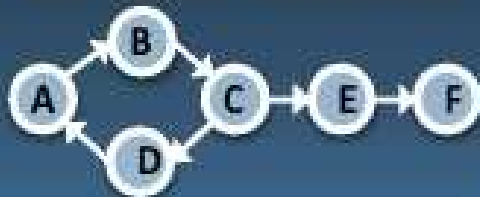
- A route between two nodes that does not use the same link (edge) twice. A-B-C-D-B is a trail, A-B-C-D-A-B is not (the link from A to B is used twice)

Path



- A route between two nodes that does not pass through the same node twice. A-B-C-D is a path, A-B-C-D-A-B is not. Only exception: closed paths – A-B-C-D-A

Directed Versions



- Walks, trails and paths have the same definition, except now direction matters. If a route includes a link in the wrong direction: semi-path, semi-trail, semi-walk.

A Closer Look at Nodes

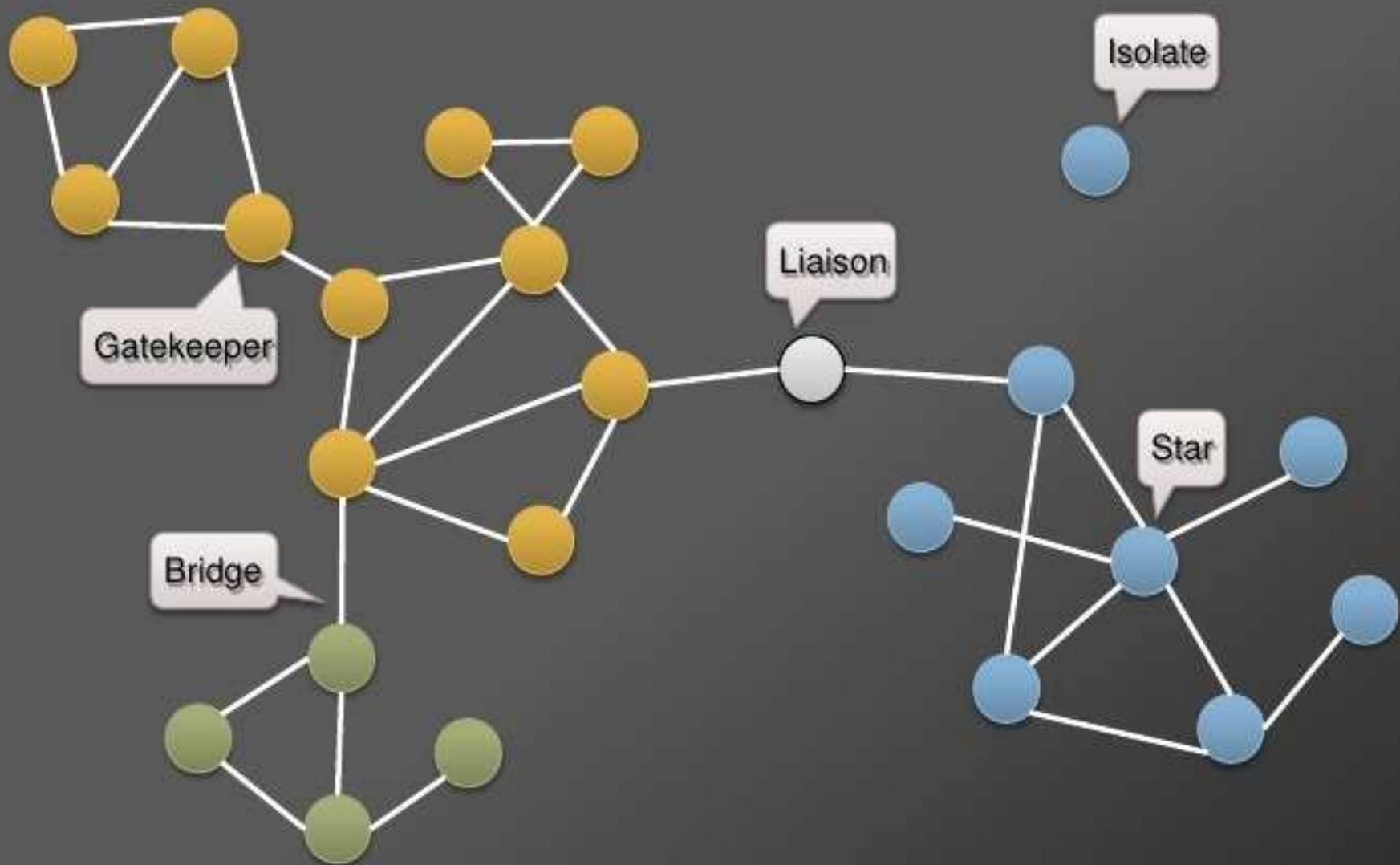
Attributes

- **Every node in a network can be assigned a range of attributes:**
 - Binary (e.g. 1 if the person has Internet access, 0 for those who are not online)
 - Categorical (e.g. 1 for Democrats, 2 for Republicans, 3 for undecided, etc.)
 - Continuous (e.g. age, income, etc.)

Roles

- **Star** - An actor who is highly central (has many connections)
- **Liaison** - An actor who has links to two or more groups that would otherwise not be linked, but is not a member of either group.
- **Bridge** - An actor who links/belongs to two or more groups (strict definitions may require that no other link between the two groups exists)
- **Gatekeeper** - An actor who mediates or controls the flow (is the single link) between one part of the network and another
- **Isolate** - An actor who has no links to other actors

Node Types



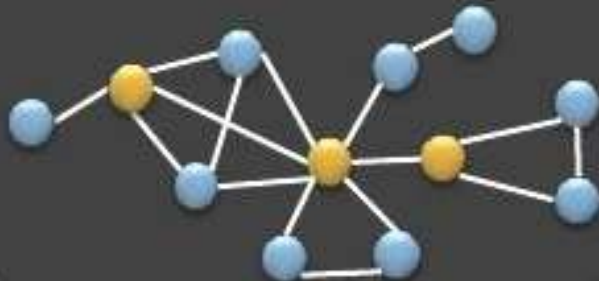
Node Centrality

Degree



- **Degree:** Number of ties a node has (also called nominations)
In a directed network: in-degree and out-degree
In-degree (number incoming ties) also called *prestige*.
Normalized: divide by total possible number of ties a single node can have ($N-1$, where N is the number of nodes in the net)
Shows how well connected the node is to the network.

Closeness



- **Closeness:** based on the average distance from a node to every other reachable node in the network. Calculated as the inverted sum of the shortest paths between the node and every other node. The smallest possible sum (star linked to every node): $N-1$
Indicates how quickly an actor can reach everyone in the network – to spread information, etc.

Betweenness



- **Betweenness:** based on the number of cases in which a node lies on the shortest path between two other nodes in the network. Adjusted by total number of shortest paths.
Tells you whether a node is in a strategic position (gatekeeper, liaison) controlling the spread of information/resources.

Social Network Measures & Mechanisms

Reciprocity



- **Reciprocity/mutuality:** tie A to B and B to A
Can be calculated as a network measure: proportion reciprocal ties from the total number of ties.
Reciprocity is calculated for directed networks - in undirected ones all links are by definition reciprocated.

Transitivity



- **Transitivity:** “the friends of my friends are my friends” principle. People look for a balanced environment and try to reduce intransitive triads to avoid dissonance.
- **Clustering:** a measure linked to the number of closed triangles in a network. Higher clustering means more “cliquishness”

Homophily



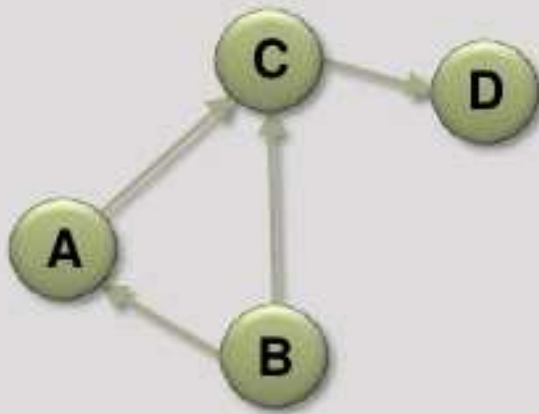
- **Homophily** is not a network measure – rather it is a mechanism that may motivate link creation and dissolution. The idea behind homophily is that people who are similar in some way (gender, political leanings, etc.) are more likely to be connected. This may be because we create links to people who are like us – or because of social influence (our friends become similar to us with time).

What does network data look like?

Although there are other options (edge lists, node lists, etc.) network data is typically used in the form of a matrix.

Row X column Y is 1 if there is a link from X to Y - and 0 if there is no link.

The diagonal represents self-loops: links from X back to X.



	A	B	C	D
A	0	0	1	0
B	1	0	1	0
C	0	0	0	1
D	0	0	0	0

How do we determine who is “important” in a Network?

The Key Player Problem

- Network Disruption problem
 - How to maximally disrupt the functioning of a network by intervening with the key players• e.g., removing them
- Network Influence problem
 - How to maximally spread ideas, misinformation, materials, diseases, etc. by seeding key players

Applications

DISRUPTION

- Who to immunize or quarantine in order to slow the spread of infectious disease?
- Who to arrest to disrupt criminal network?
- Where is an organization most vulnerable to turnover?

INFLUENCE

- **Selecting peer advocates for diffusing safe practices**
- **Who to “turn” or feed false information to?**
- **Select subset of employees for intervention prior to change initiative**

Tracking Two Identified Terrorists

Early in 2000, the CIA was informed of two terrorist suspects linked to al-Qaeda.

Nawaf Alhazmi and **Khalid Almihdhar** were photographed attending a meeting of known terrorists in Malaysia

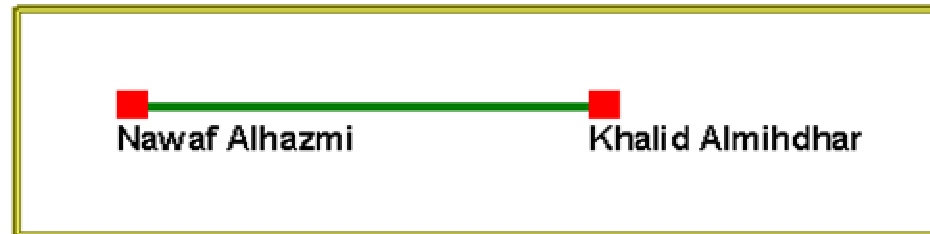


Figure 1 - Two known suspects in January 2000

**Arrest or deport these suspects immediately?
No, intelligence services need to discover more of the al-Qaeda network. Once suspects have been discovered, we can use their daily activities to uncloak their network.**



who they call / email



who visits with them locally and in other cities



where their money comes from



picture of the terrorist organization

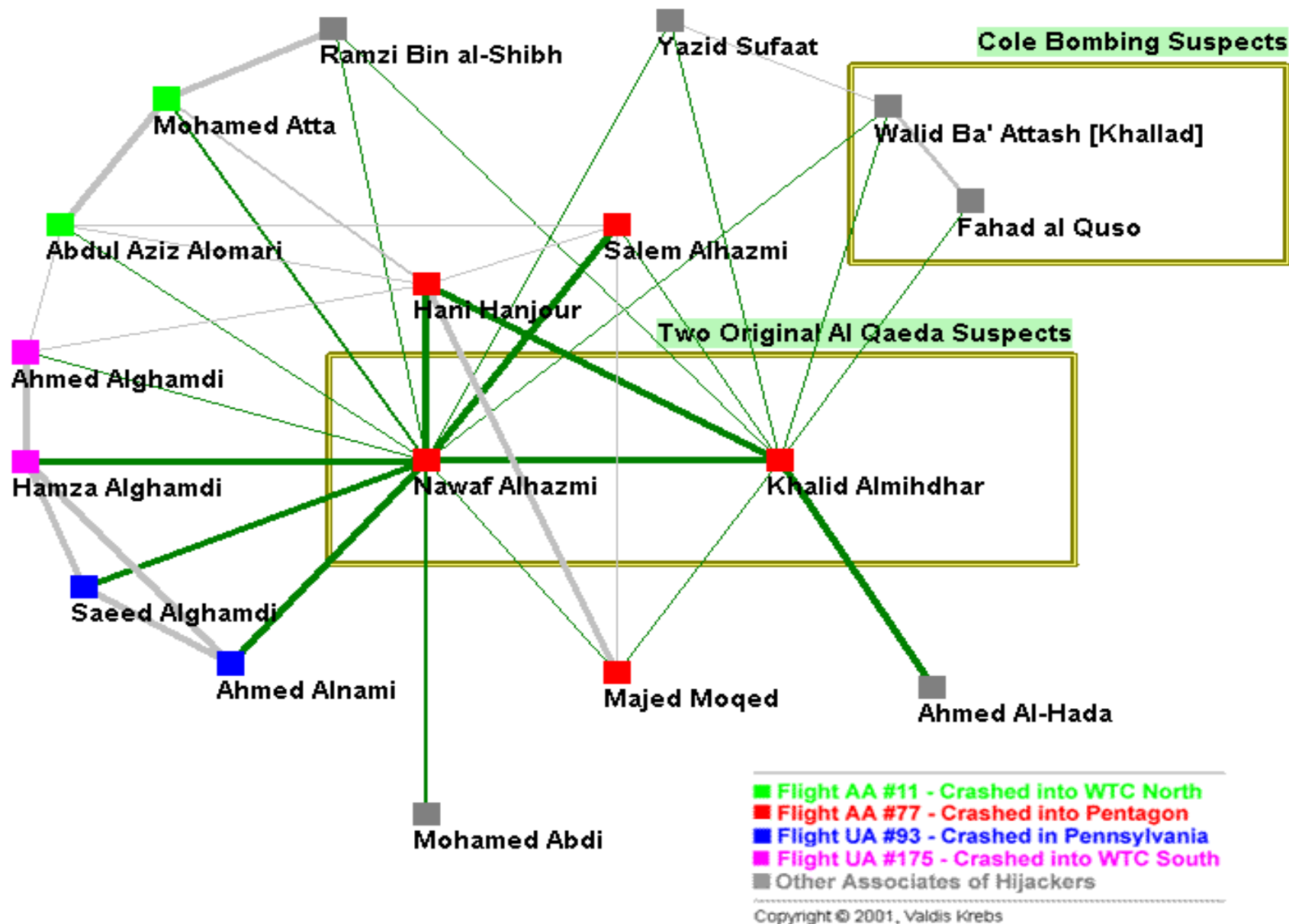


Figure 2 - All nodes within 1 step [direct link] of original suspects

The figure shows the two suspects and their immediate ties

- direct ties are colored green
- link thickness show the strength of connection



their indirect ties (the “connections of their connections”)

Atta is just another node to be investigated.

The investigator unclocking Mohammed Atta's network would not be aware of Atta's eventual importance.

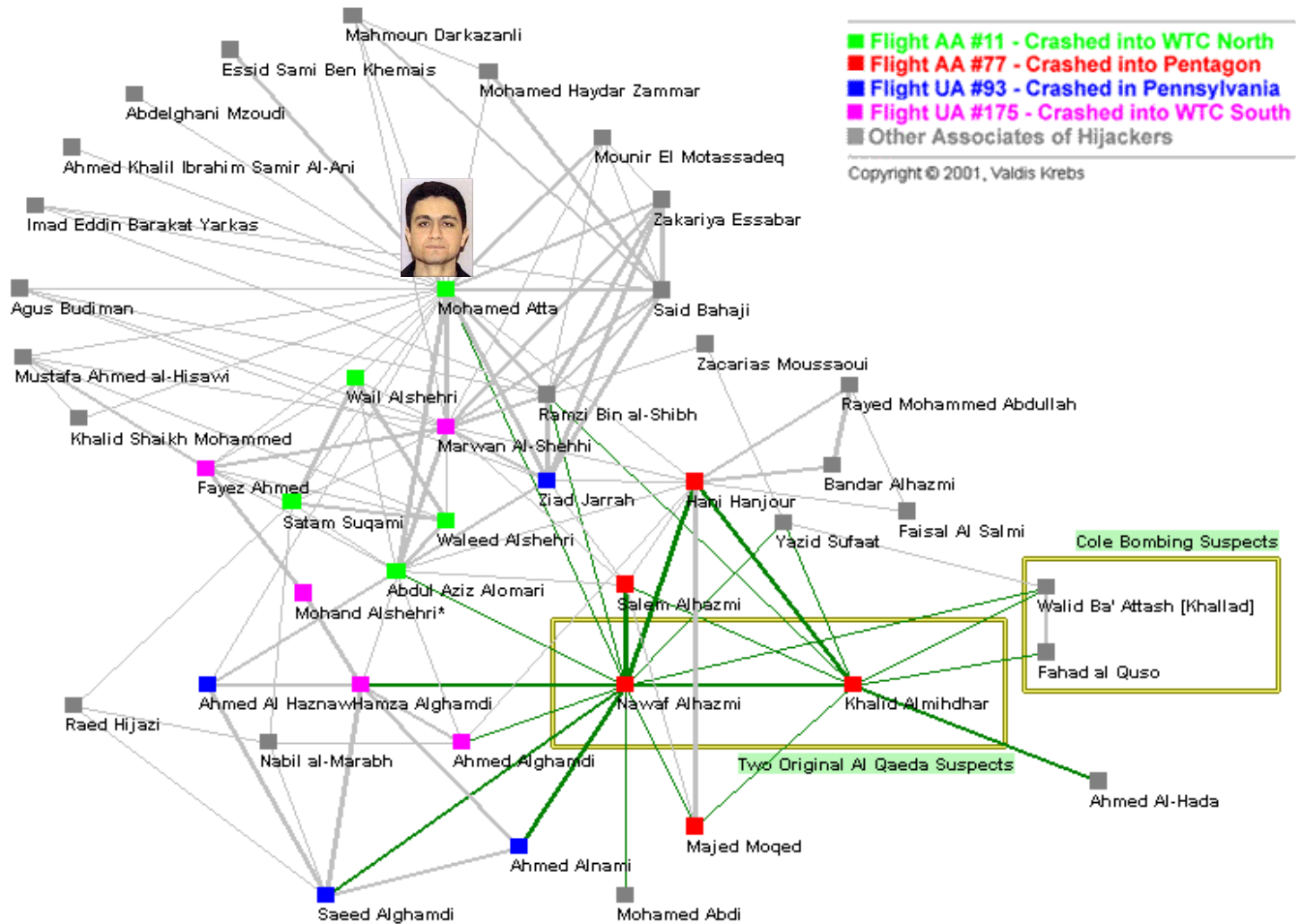


Figure 3 - All Nodes within 2 steps / degrees of original suspects


the direct connections of the original suspect 
their indirect connections 

Social network metrics reveal



Mohammed *Att*a *em*erging as the local



Mohamed Atta, an 
Egyptian national, was the
ringleader of the hijackers.

Big Data Analytics

- Examining large amount of data
- Appropriate information
- Identification of hidden patterns, unknown correlations
- Competitive advantage
- Better business decisions: strategic and operational
- Effective marketing, customer satisfaction, increased revenue

Resources:

<http://www.orgnet.com/prevent.html>

http://www.slideshare.net/kateto/network-analysis-basics-and-applications-to-online-networks?qid=892b01a1-19a9-459f-87e1-61a356f84f98&v=qf1&b=&from_search=12

On-Line UCINET:

www.analytictech.com

This is the company website.

<http://faculty.ucr.edu/~hanneman/nettext/>

This is an on-line reference manual to UCINET.

<http://tech.groups.yahoo.com/group/ucinet>

An online community of practice for UCINET.

