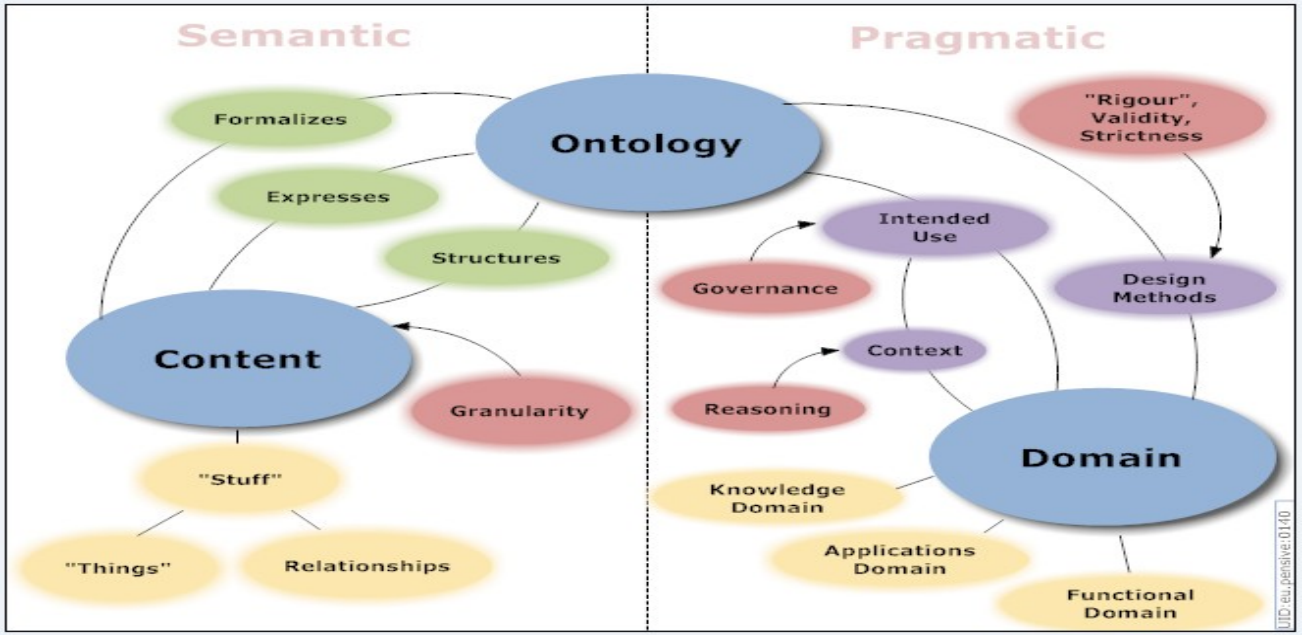
Sémantický web, ontologie. Sociální sítě.



Sémantický web

- Metody a techniky pro přiřazení významu (sémantiky) informacím na webu
- Web rozšířený o metadata
- **Metadata** = data o datech
- Postaven na formátu **RDF**

Cíle sémantického webu

- **Integrovat data** z různých zdrojů
- Umožnit **výměnu dat** mezi aplikacemi napříč celým webem
- Umožnit **kvalitnější strojové vyhledávání** informací na webu
- Umožnit **popsat vztahy** mezi daty a objekty v reálném světě
- **Přiřadit** informacím na webu přesný **význam**

Metadata v HTML

- Pomocí **<meta>** tagů:
  ```
  <meta name="keywords" content="HTML, CSS, XML" />
  ```
- Cíl: umožnit kvalitnější vyhledávání, než obyčejný full-text search
- Zneužíváno ve velké míře spammery
- Neumožňuje definovat vztahy a hierarchie objektů
- Dnes vyhledávače dávají přednost jiným metodám, než prohledávání <meta> tagů

RDF

- **RDF** = Resource Description Framework
- Framework pro popis zdrojů na webu
- Navržen tak, aby byl strojově čitelný a pochopitelný
- Doporučení W3C
- Různé způsoby serializace (uložení do souboru), př. **RDF/XML**

Princip RDF

- Každému zdroji na webu přiřadí trojici:
  - Subject (subjekt, podmět)
  - Predicate (predikát, vlastnost)
  - Object (objekt, předmět)
- Při definici subjektů a predikátů je typicky potřeba definovat **URI** (Unique Resource Identifier) pro jednoznačné přiřazení významu.
- RDF dokumenty lze ukládat do **triplestore** databází (databáze optimalizované pro RDF trojice) nebo serializovat pomocí XML (formát **RDF/XML**)

RDF/XML

- Příklad: „Obloha má modrou barvu.“
  - Podmět: „obloha“
  - Vlastnost: „mít barvu“
  - Předmět: „modrá“ („blue“)
- Serializace ve formátu RDF/XML:

```
1: <?xml version="1.0"?>
2:
3: <rdf:RDF
4:     xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
5:     xmlns:sky="http://fi.muni.cz/rdf/sky/">
6:     <rdf:Description rdf:about="http://fi.muni.cz/rdf/sky">
7:         <sky:color>blue</sky:color>
8:     </rdf:Description>
9: </rdf:RDF>
```

Triplestores

- Databáze optimalizované pro ukládání RDF trojic (subjekt, predikát, objekt)
- Mnoho implementací v různých jazycích (C, C#, PHP, Java, Perl)
- Postaveny buď nad existujícím relačním databázovým strojem (MySQL, PostgreSQL, MS SQL, Oracle), nebo vyvinuty kompletně od začátku přesně pro svůj účel (vyšší efektivita)

Ontologie

- Model pro popis světa složeného z typů, vlastností a vztahů
- Využití v sémantickém webu pro přiřazení významu datům (tj. pro tvorbu metadatového modelu)
- Při tvorbě ontologií je snaha o co nejpřesnější podobnost mezi objekty reálného světa a vlastnostmi modelu

Kategorie ontologií

- **Individua** (instance a objekty)
- **Třídy** (množiny, kolekce, pojmy, typy, druhy)
- **Atributy** (aspekty, stavy, vlastnosti, charakteristiky a parametry, kterých mohou objekty/třídy nabývat)
- **Relace** (způsoby, jakými k sobě mohou třídy a individua navzájem patřit)
- **Funkční výrazy** (komplexní struktury nad relacemi)

Kategorie ontologií

- **Restrikce** (formální popis platného vstupu)
- **Pravidla** (Příkazy ve formě if-then (příčina-následek) popisující logické inference, které mohou být odvozeny z výroků v dané formě)
- **Axiomy** (výroky (vč. pravidel) v logické formě, které dohromady skládají kompletní teorii, kterou ontologie popisuje. Nemusí obsahovat pouze apriorní znalosti, ale také odvozené teorie z jiných axiomů.
- **Události** (změny atributů a relací)

Inference znalostí

- Pojem **inference**
  - 1) dobře navržená logická heuristika pro odvozování nových znalostí
  - 2) odvozená znalost
- **Inference znalostí** - odvozování nových znalostí na základě existujících (známých) znalostí (inferencí)
- Využití v sémantickém webu při **strojovém vyhledávání** nových znalostí

Inferenční enginy

- Počítačové programy, které zkouší odvodit odpověď **z báze znalostí** (knowledge base, množina axiomů/výroků/faktů/znalostí/popř. inferencí)
- Data v bázi znalostí musí být uložena takovým způsobem, aby stroj/engine dokázal odvodit a porozumět jejich významu, tj. musí být explicitně vyjádřena jejich **sémantika** (samotná data musí být doplněna o **metadata**)

SPARQL [„spa:kl“]

- Jazyk / protokol pro inferenci znalostí z RDF dokumentů
- Umožňuje provádět dotazy nad RDF trojicemi (triplestore databázemi)
- Podobná syntax jako SQL
- Výhoda SPARQL: dotazy jsou díky přítomnosti URI v RDF formátu globálně jednoznačné

Sociální sítě

- propojená skupina lidí, kteří se navzájem ovlivňují
- **Sociální software** (**socioware**) - software, který umožňuje tvořit komunity pomocí počítačových propojení.
- **Virtuální komunita**, **e-komunita**

Periferní (tj. lurker – *číhající*) - externí, nestrukturovaná účast

Příchozí (tj. nováček) – nově příchozí je vpuštěn do komunity a může se plně účastnit diskuze

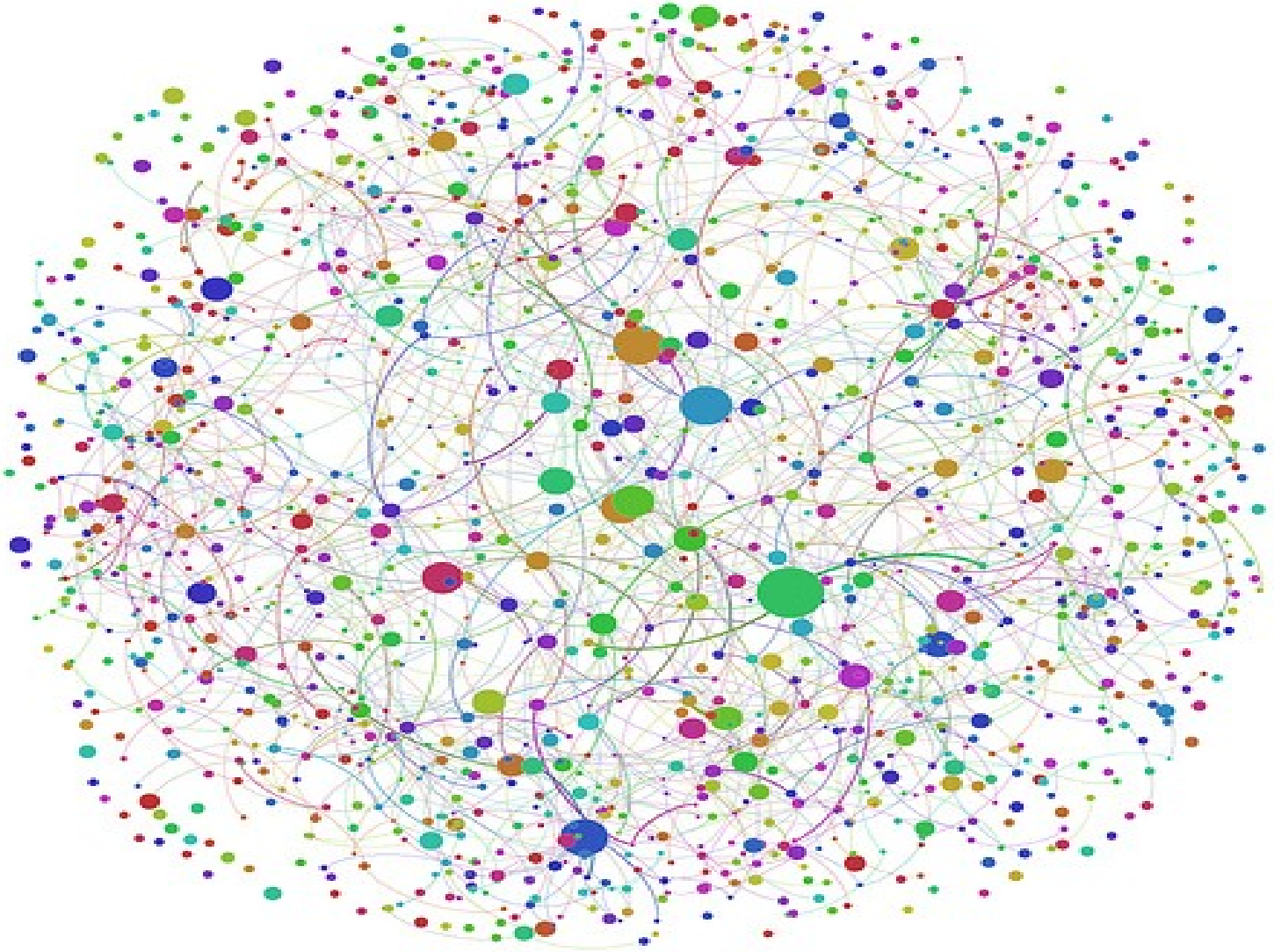
Zasvěcenec (tj. stálý člen) – plně uznaný účastník

Strážce hranic (tj. vůdce) – podporuje členství a zprostředkovává interakce

Odchozí (tj. starý) – proces opouštění komunity kvůli novým vztahům, novým místům, novým vyhlídkám

Sociální sítě

- **Facebook**
- **Twitter** (tweety, „SMS Internetu“)
- **MySpace** – sdílení hudby a videa
- **Orkut –** sdílení multimédií, chatování a hledání ztracených přátel.
- **Classmates (**Spolužáci.cz)
- **Blackplanet -** síť určená pro Afroameričany a jejich přátele
- **Hi5, Friendster, Bebo, …**



Data: AER, JPE, Econometrica, RES, QJE (2000-present)
By Cloudly From: www.cloudlychen.net

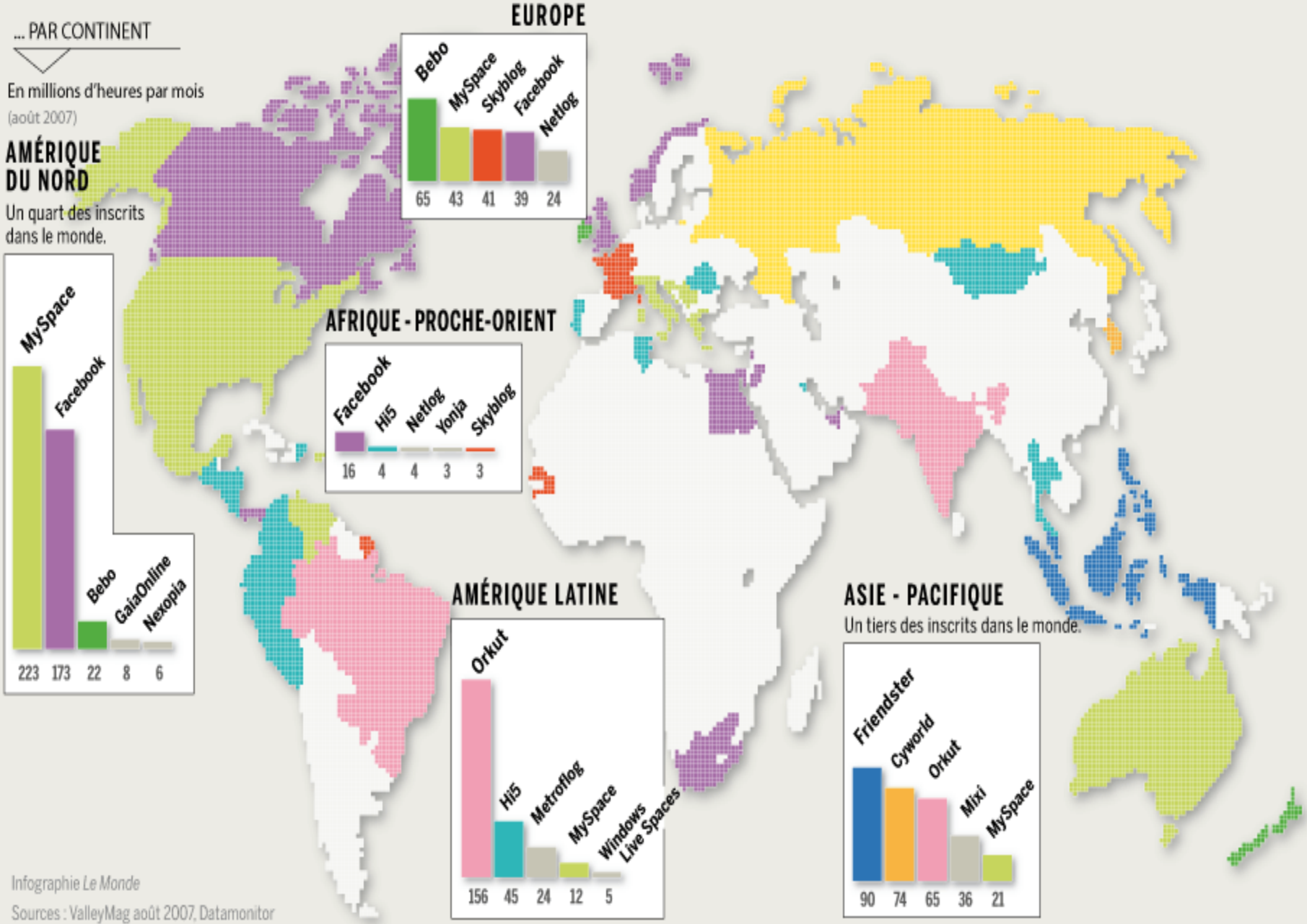
LES SITES DE RÉSEAUX SOCIAUX LES PLUS FRÉQUENTÉS...

... PAR PAYS

Nom du site	MySpace	Facebook	Bebo	Cyworld	Skyblog	Hi5	Friendster	Orkut	Live Journal
Nationalité de l'entreprise :	Etats-Unis	Etats-Unis	Etats-Unis	Corée du Sud	France	Etats-Unis	Etats-Unis	Etats-Unis	Russie

... PAR CONTINENT

En millions d'heures par mois
(août 2007)



Infographie *Le Monde*

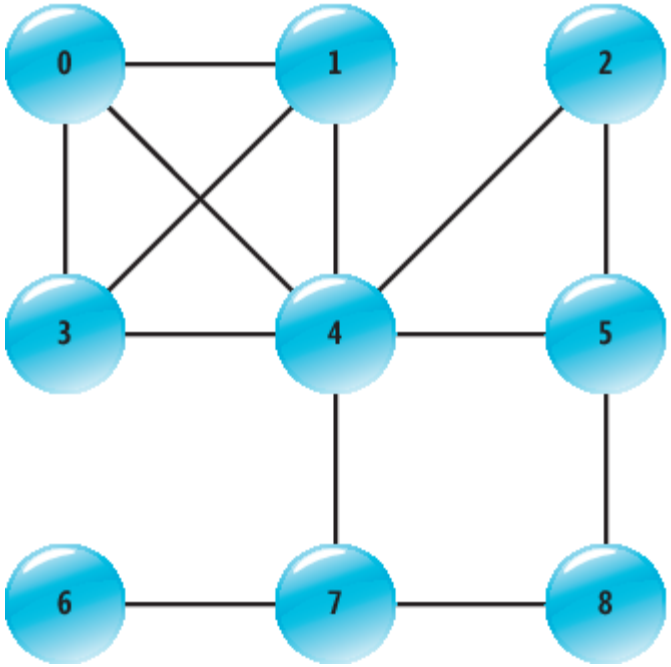
Sources : ValleyMag août 2007, Datamonitor

Modelování a analýza sociálních sítí

- Grafy
- Matice
- Vizualizace
- Aplikace

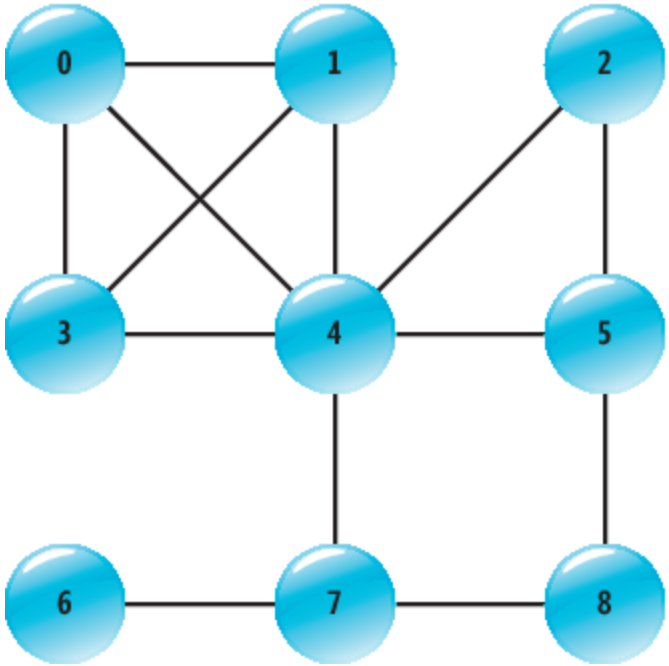
Graf

- **Jednoduchý meorientovaný graf** je dvojice G = (V, E) , kde V je neprázdná množina **vrcholů** (**uzlů**) a E je množina dvouprvkových množin vrcholů, tzv. **(neorientovaných) hran**.
- **Jednoduchý orientovaný graf** je dvojice G = (V, E) , kde V je neprázdná množina vrcholů (uzlů) a E je množina uspořádaných dvojic vrcholů, tzv. **(orientovaných) hran**.


0
1
2
3
4
5
6
7
8

Maticová reprezentace grafu

	0	1	2	3	4	5	6	7	8
0	0	1	0	1	1	0	0	0	0
1	1	0	0	1	1	0	0	0	0
2	0	0	0	0	1	1	0	0	0
3	1	1	0	0	1	0	0	0	0
4	1	1	1	1	0	1	0	1	0
5	0	0	1	0	1	0	0	0	1
6	0	0	0	0	0	0	0	1	0
7	0	0	0	0	1	0	1	0	1
8	0	0	0	0	0	1	0	1	0

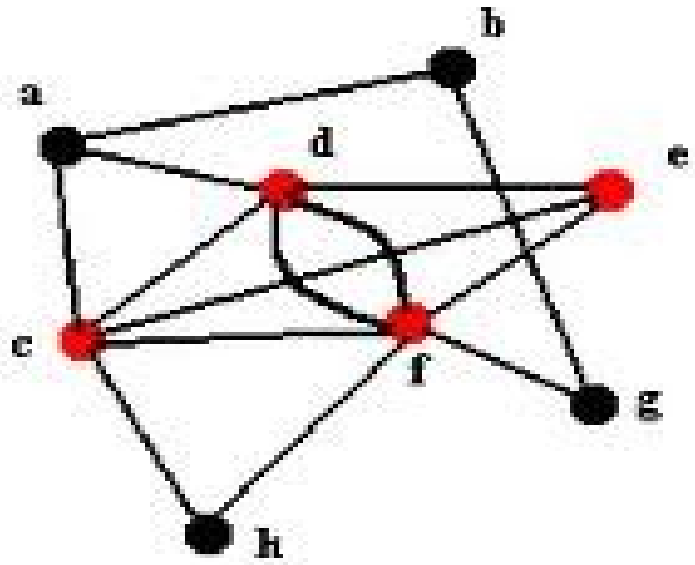


	0	1	2	3	4	5	6	7	8
0	0	1	0	1	1	0	0	0	0
1	1	0	0	1	1	0	0	0	0
2	0	0	0	0	1	1	0	0	0
3	1	1	0	0	1	0	0	0	0
4	1	1	1	1	0	1	0	1	0
5	0	0	1	0	1	0	0	0	1
6	0	0	0	0	0	0	0	1	0
7	0	0	0	0	1	0	1	0	1
8	0	0	0	0	0	1	0	1	0

Clique - klika

- social sciences: **"clique"** popisuje skupinu cca. 2 až 12 (průměr 5 - 6) osob které spolu interagují mnohem častěji a intensivněji než ostatní
- Teorie grafů: **clique** je taková podmnožina neorientovaného grafu, ve které jsou každé dva uzly spojené hranou.

Klika - př., graf

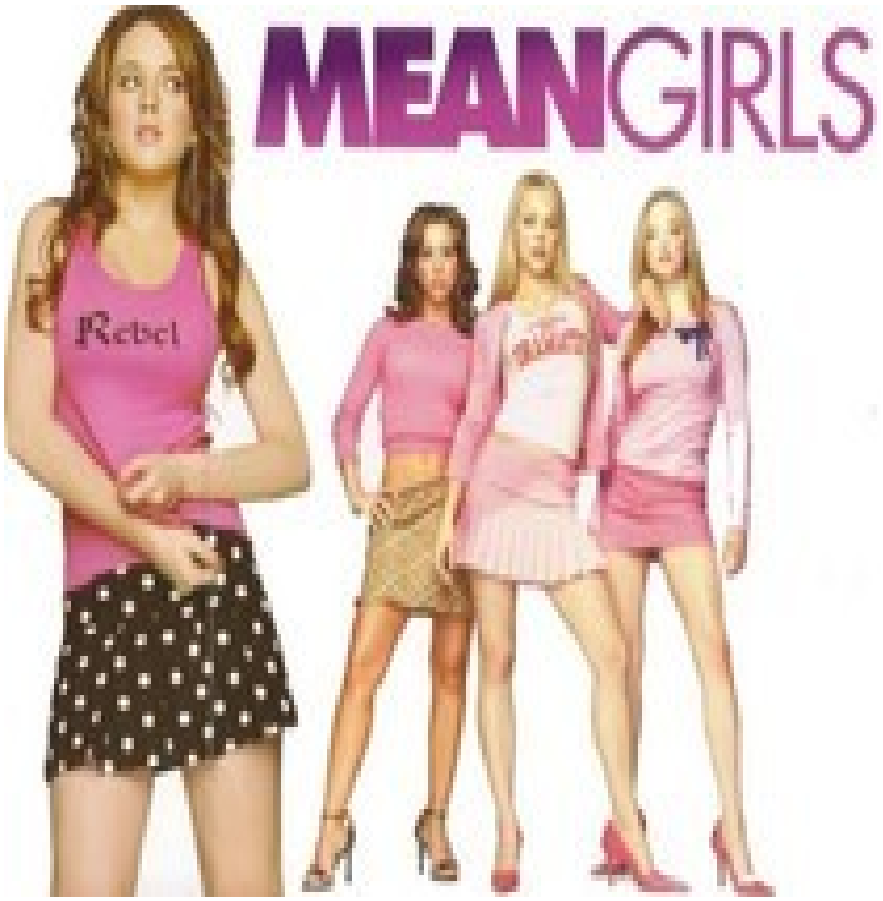


Cliques

a,b
a,c,d
b,g
c,d,e,f
c,f,h
f,g

Typy klik – př.

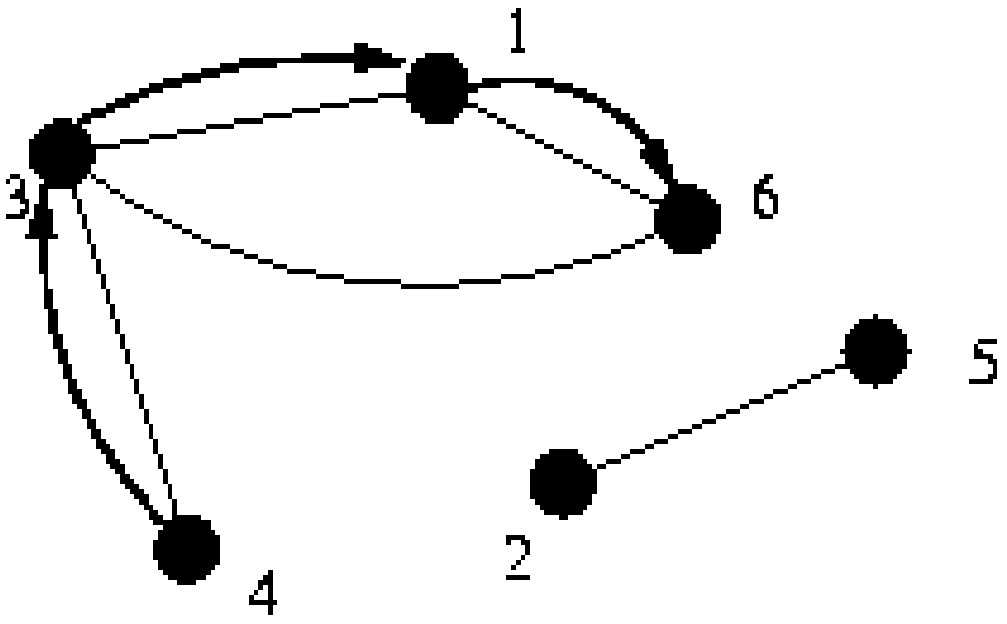
- Punkeři
- Gangsteři
- Mean girls
- Šprti (nerds)
- Skateři
- Outsideři
- Intelektuálové
- …



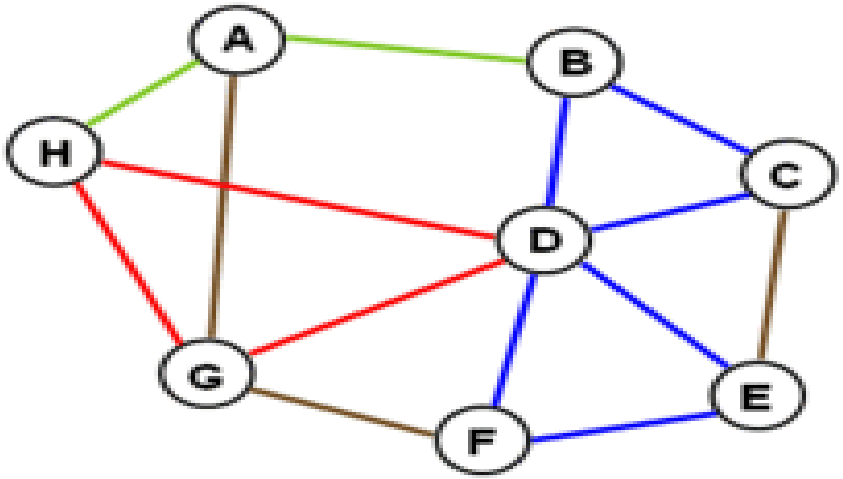
Některé další základní relevantní grafové pojmy

- Cesta
- Souvislost grafu
- Cyklus
- Strom
- Most
- Bipartitní graf
- Orientovaný graf
- Planární graf, multigraf

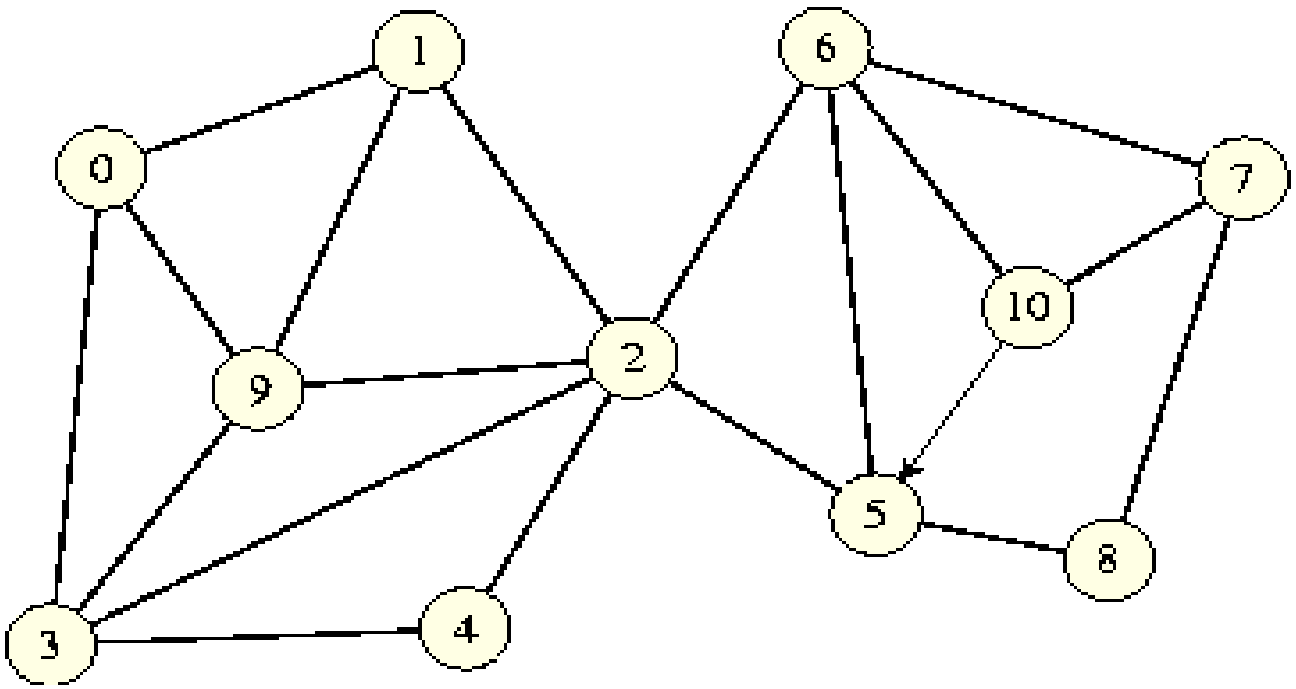
Cesta v grafu



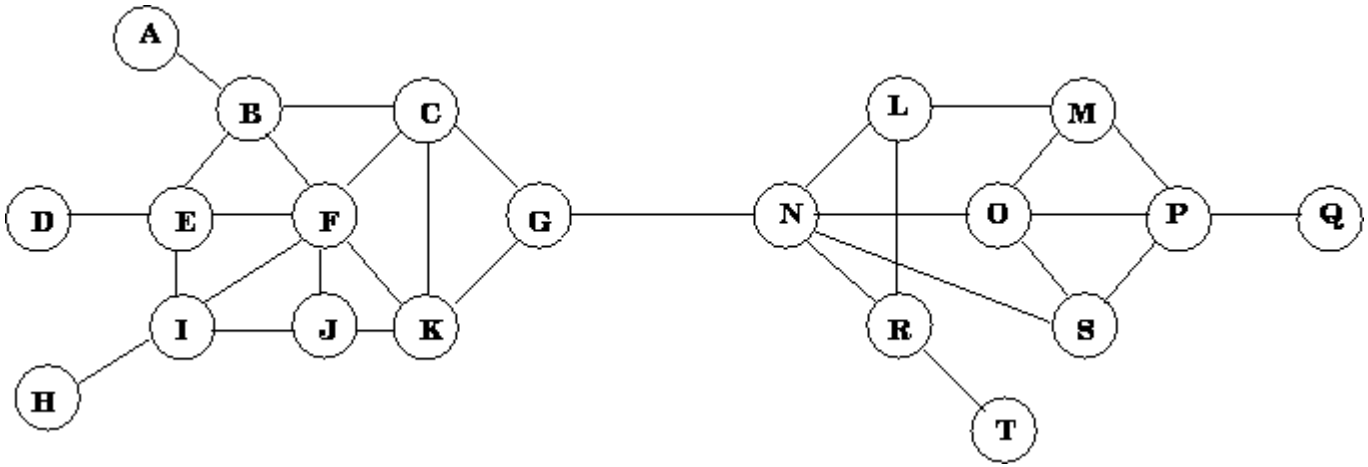
Uzavřená cesta



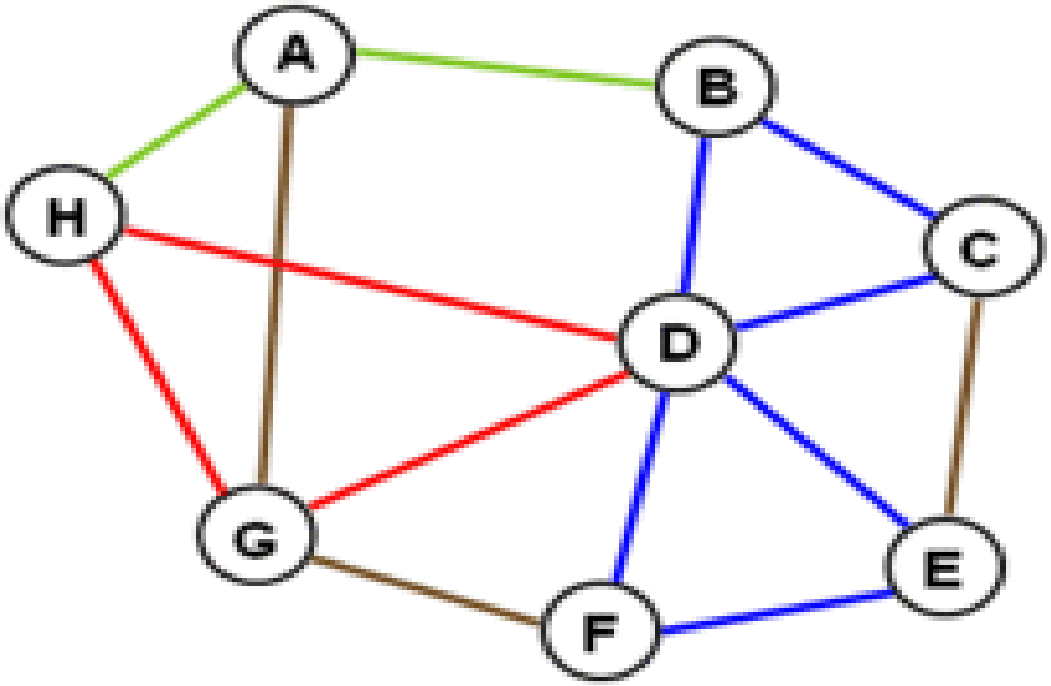
Souvislý graf



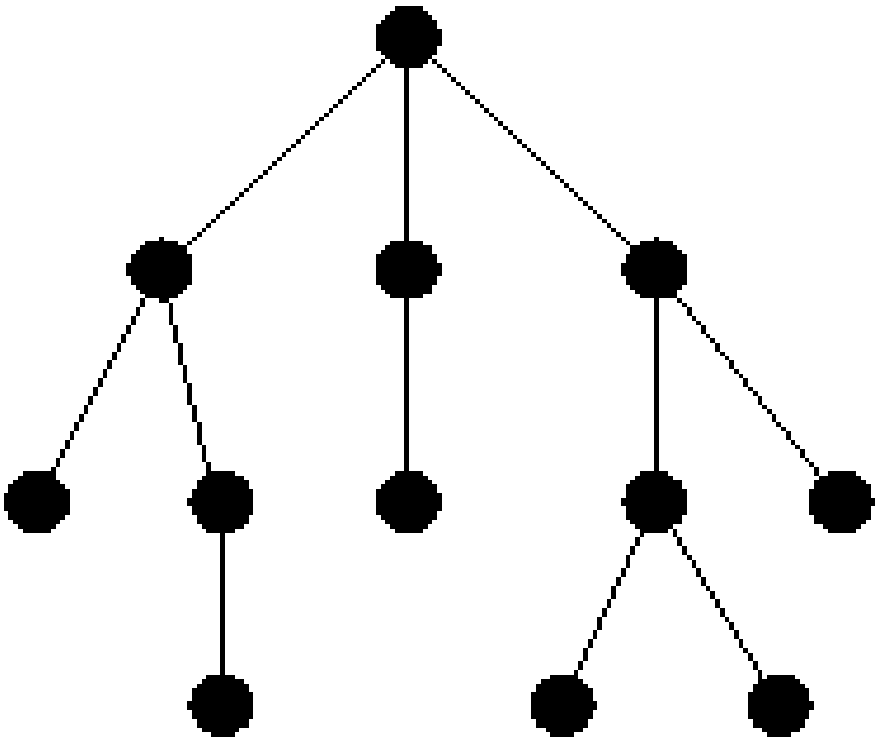
Most a bod řezu (cutpoint)



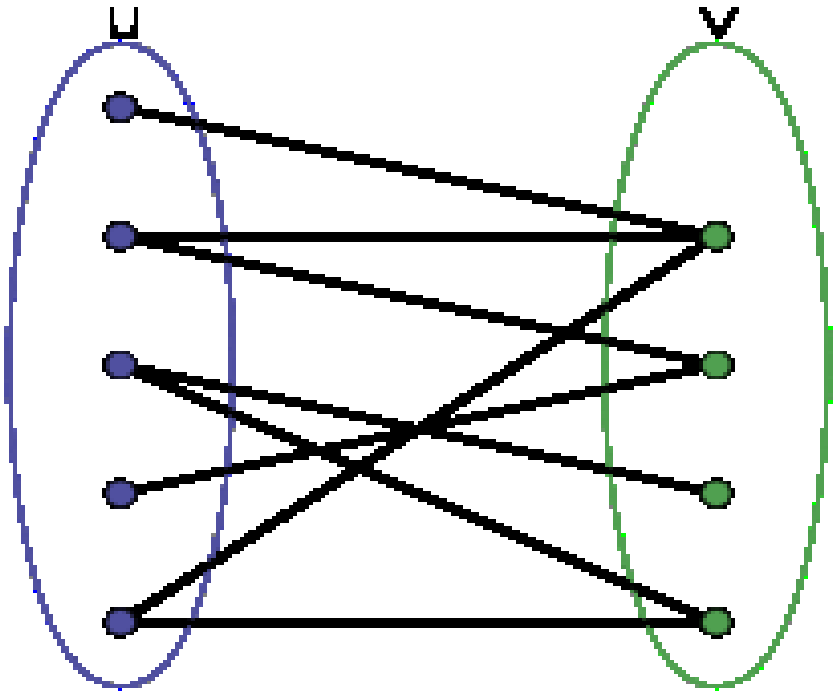
Cyklus



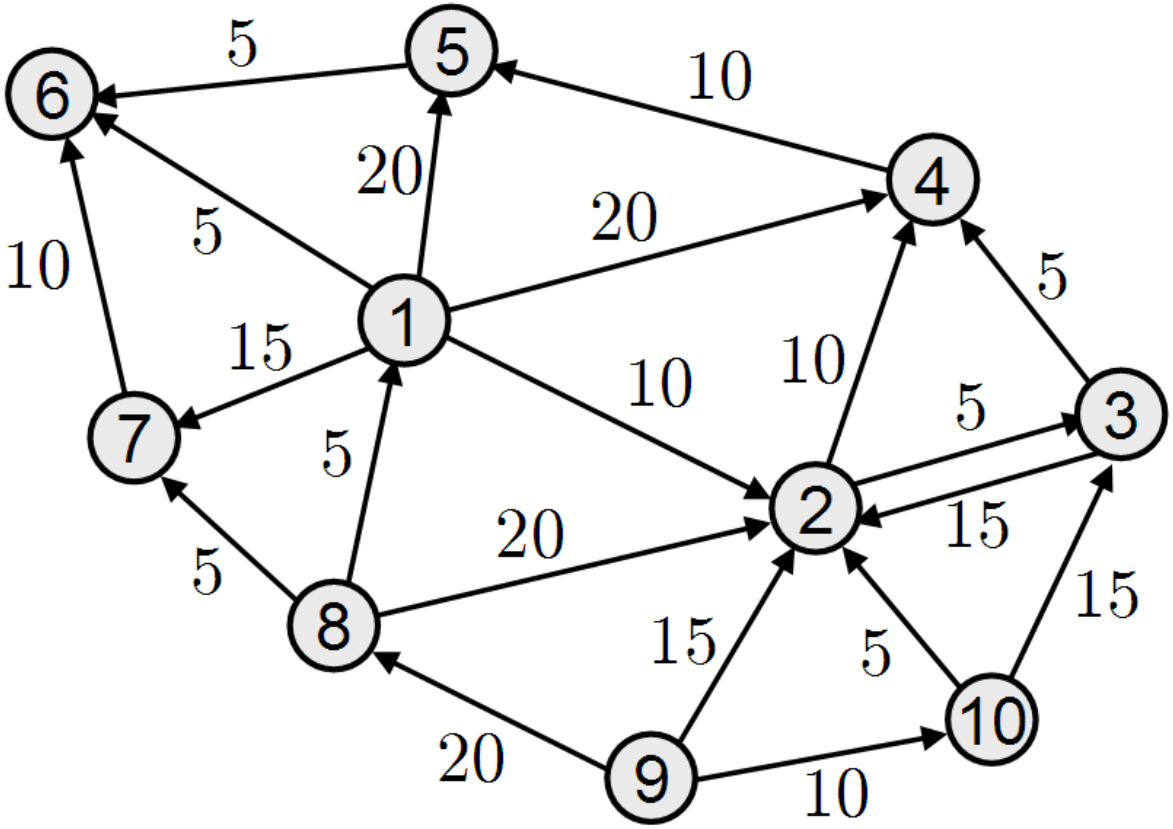
Strom



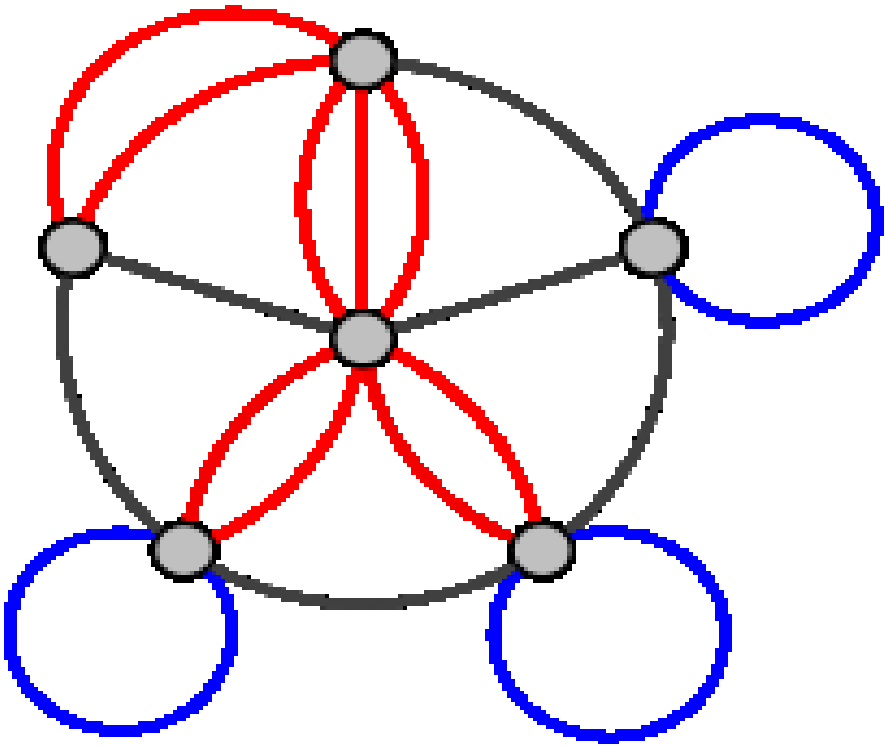
Bipartitní graf

- 

(ohodnocený) orientovaný graf



Multigraf



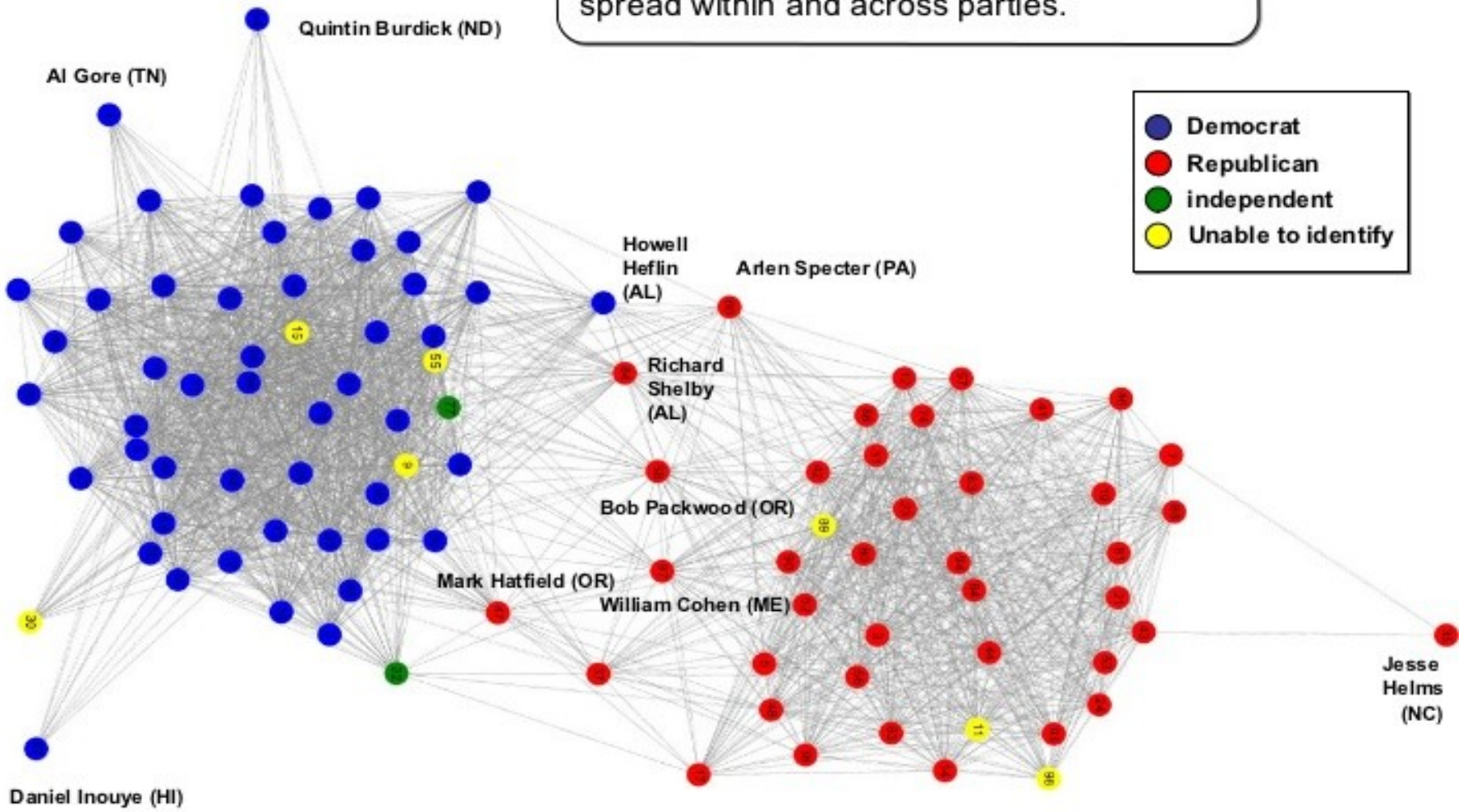
Sociální graf senátu USA

- **O'Reilly Media**
- Senatoři jsou propojeni hranou jestliže volí stejně v 65% případů během dvouletého období
- **http://www.slideshare.net/oreillymedia/us-senate-social-graph-1991-present?type=presentation**

102nd Session

January 3, 1991, to January 3, 1993

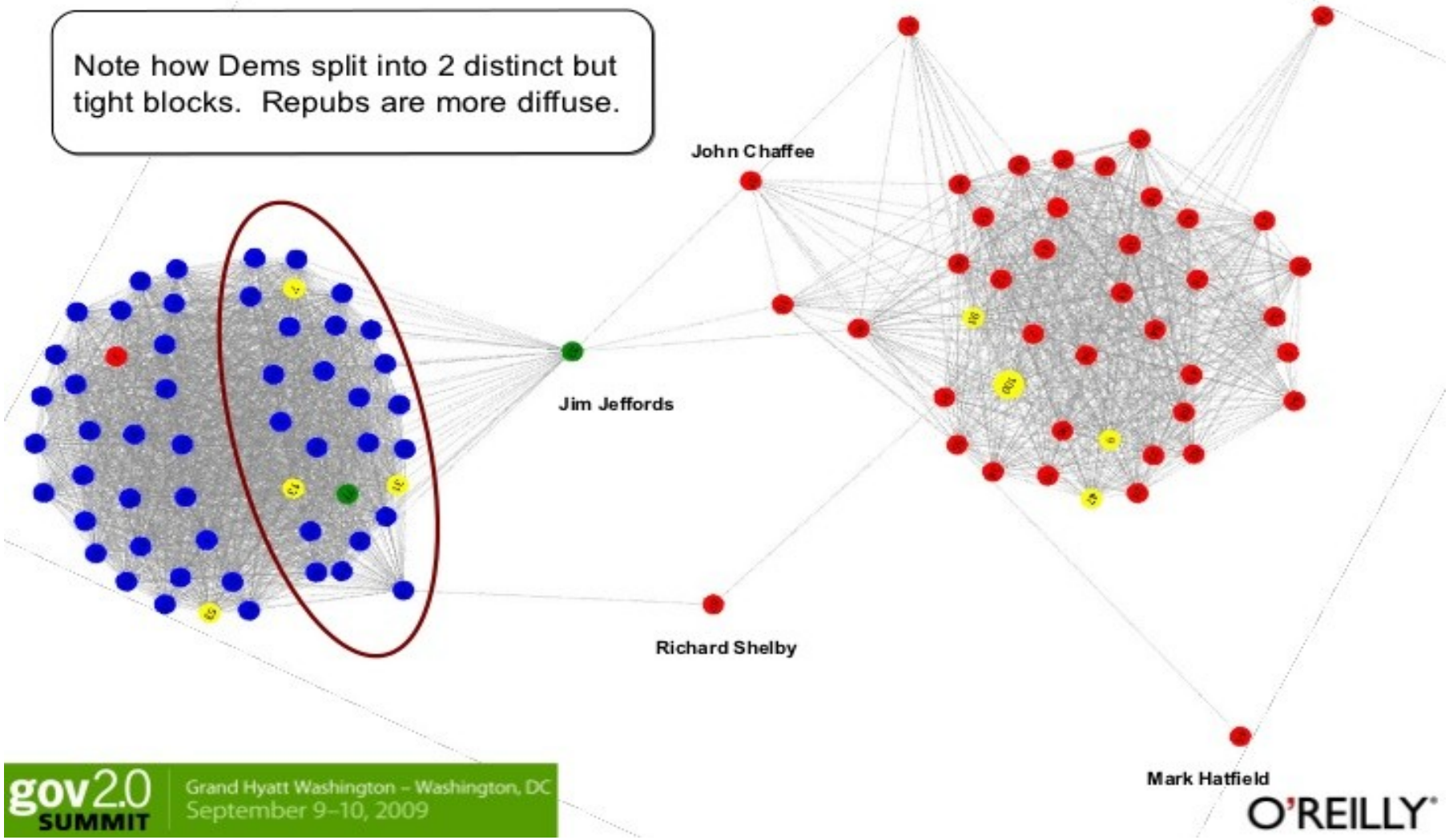
First Gulf war. Dems hold majority.
Fairly weak voting blocks with considerable spread within and across parties.



103rd Session

January 3, 1993 to January 3, 1995

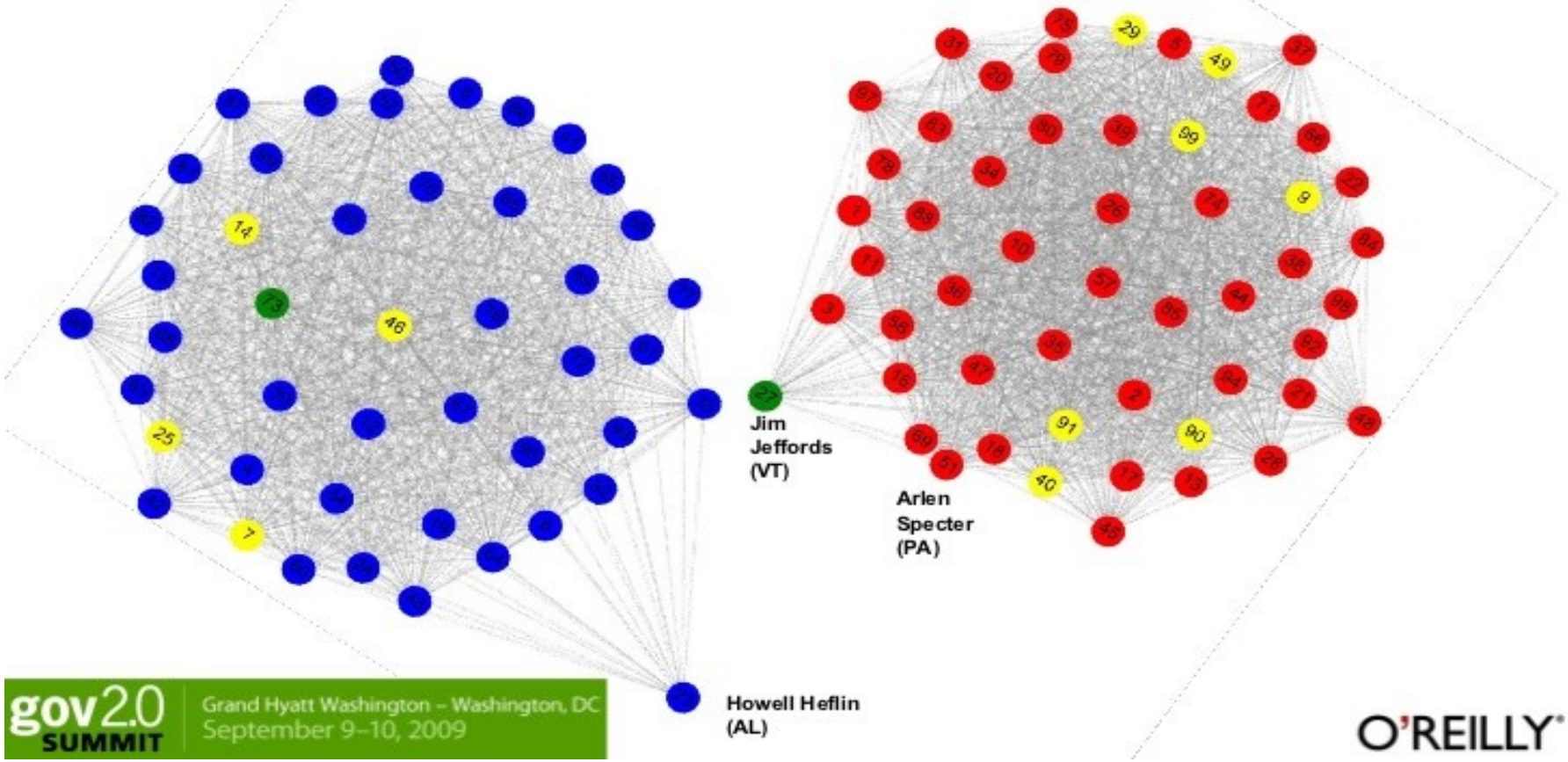
Note how Dems split into 2 distinct but tight blocks. Repubs are more diffuse.



104th Session

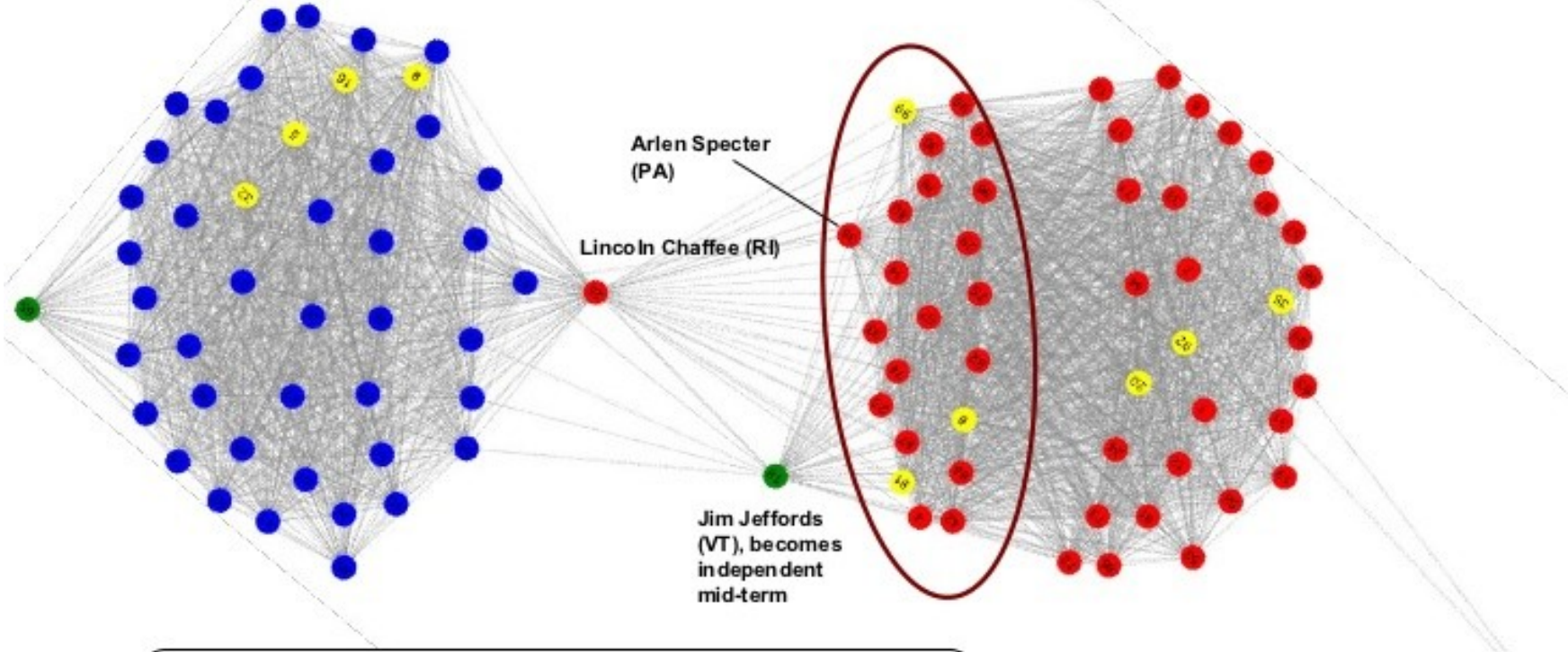
January 3, 1995 to January 3, 1997

1994 Republican Revolution. Repubs gain majority of both houses for first time since 1950s. No cross-party connections. Both parties form solid blocks.



106th Session

January 3, 1999 to January 3, 2001

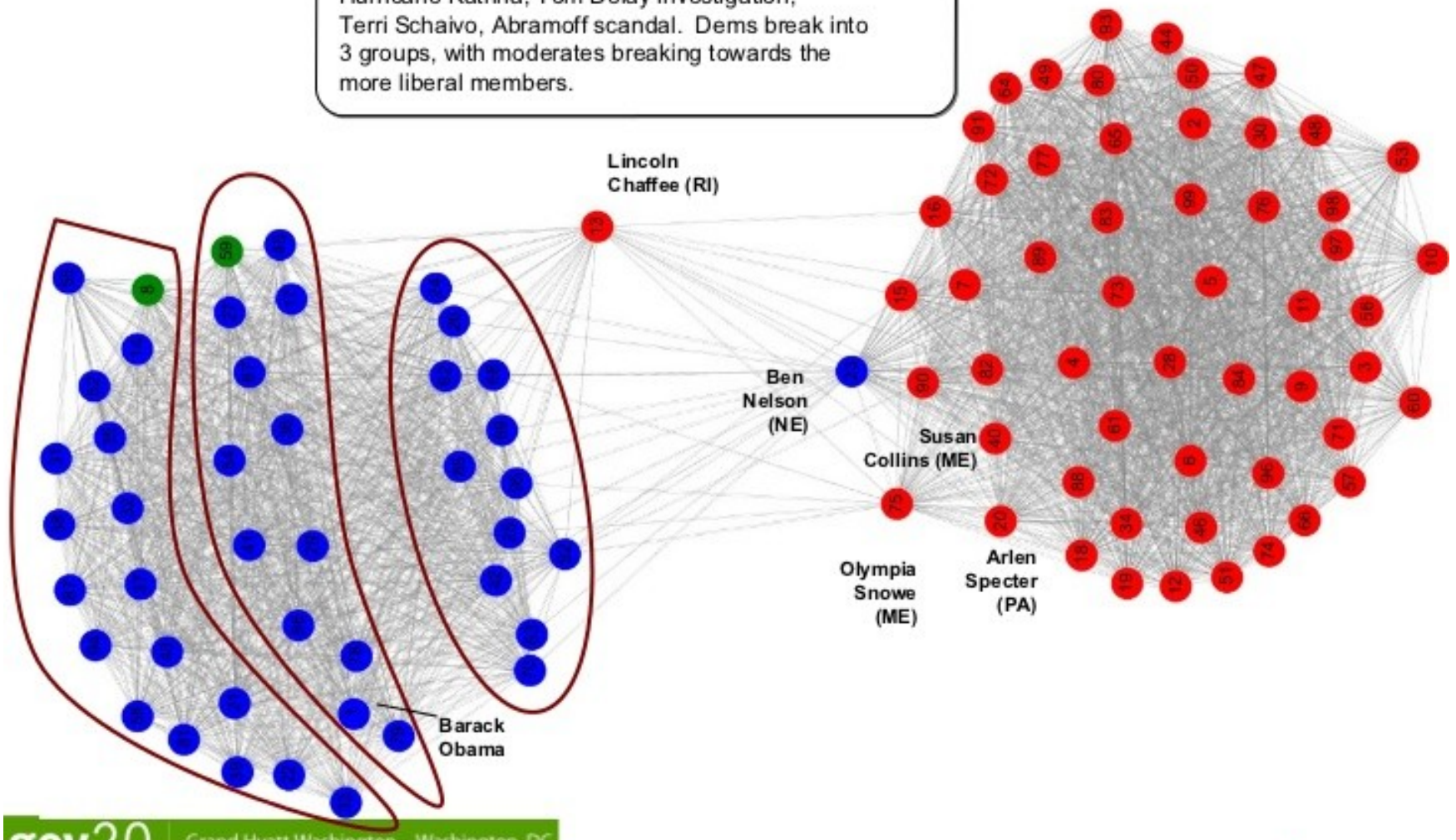


Clinton impeachment trial early in this session. Note how Dems generally maintain unity, but Repubs. fracture into 2 groups. Session ends with Bush v. Gore.

109th Session

January 3, 2005 to January 3, 2007

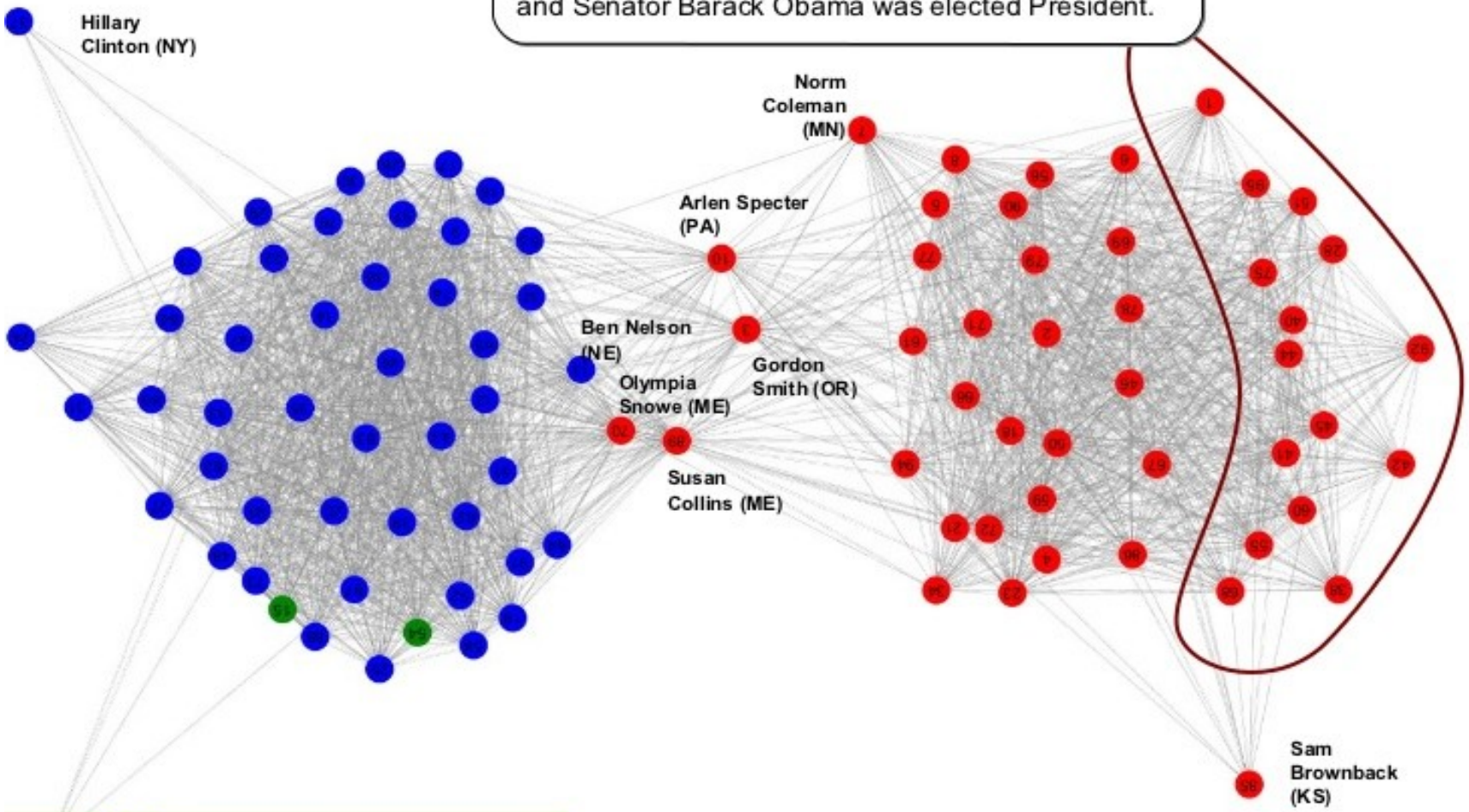
Hurricane Katrina, Tom Delay investigation, Terri Schaivo, Abramoff scandal. Dems break into 3 groups, with moderates breaking towards the more liberal members.



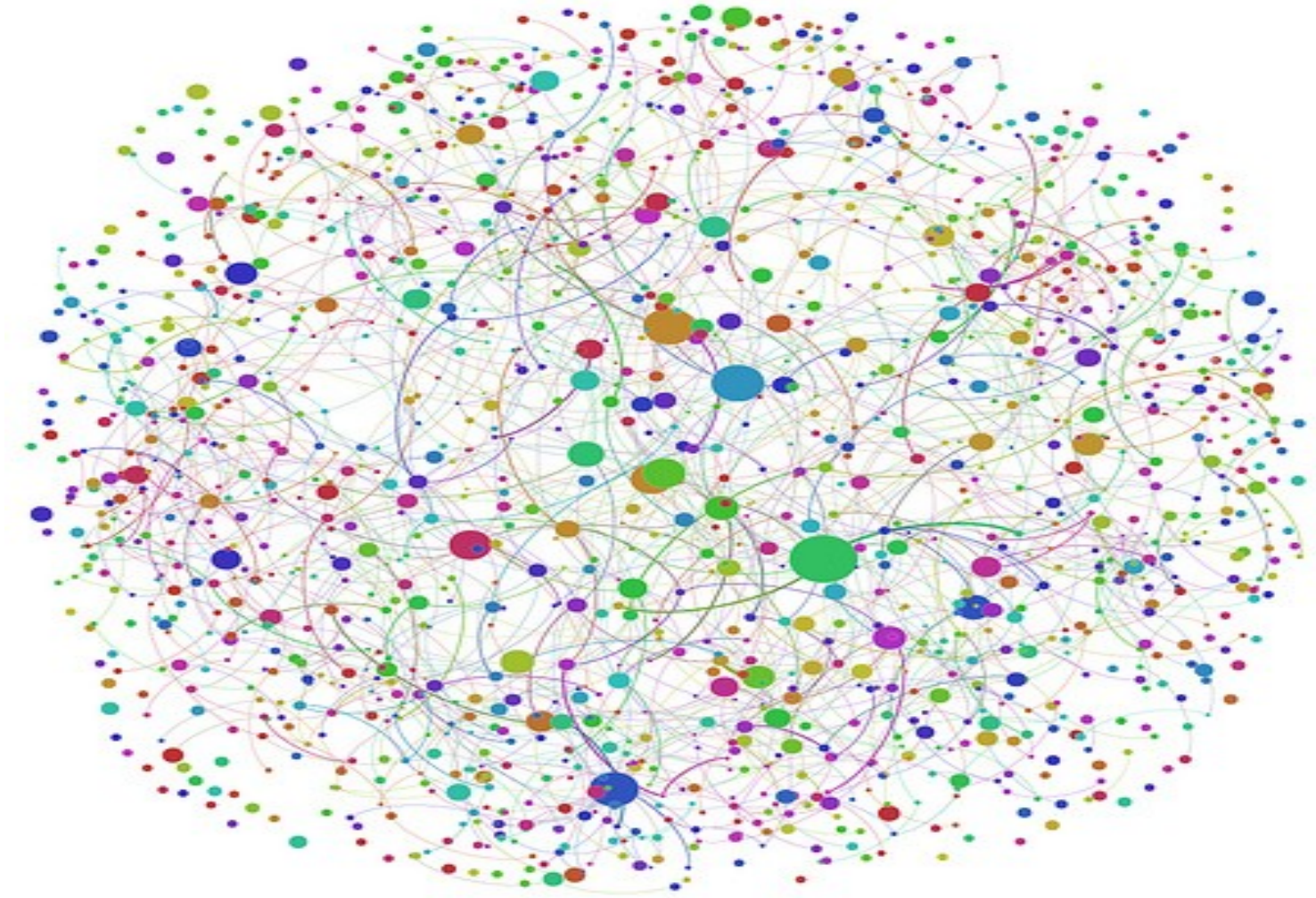
110th Session

January 3, 2007 to January 3, 2009

Conservative Repubs split into identifiable voting block. Split in Dems is reduced. Democrats increased their congressional majorities at mid-term and Senator Barack Obama was elected President.



Vizualizace



Data: AER, JPE, Econometrica, RES, QJE (2000-present)
By Cloudly From: www.cloudlychen.net

Fragment sítě Facebook