

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ

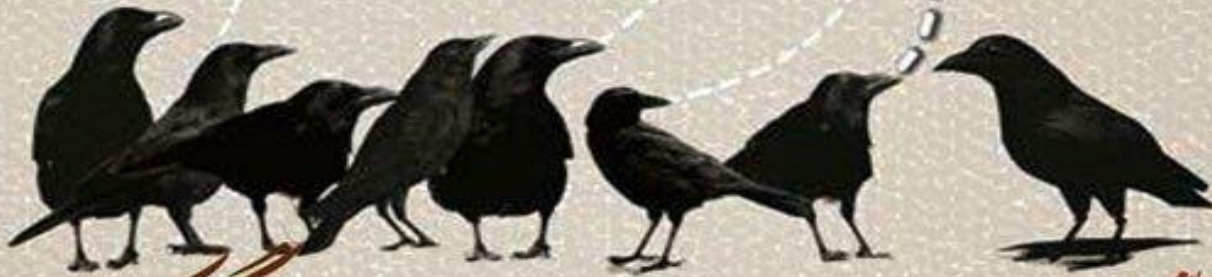
وَإِنْ تَطِيعْ أَكْثَرَ مَنْ فِي الْأَرْضِ يُضِلُّوكَ عَنْ سَبِيلِ اللَّهِ ۚ
إِنْ يَتَّبِعُونَ إِلَّا الظَّنَّ وَإِنْ هُمْ إِلَّا يَخْرُصُونَ (انعام ۱۱۶)

و اگر از اکثر مردم روی زمین پیروی کنی تو را از راه خدا گمراه خواهند کرد، که اینان جز از هوا و هوس نمی‌روند و جز دروغ چیزی در دست ندارند.

اطلاعات به معنای دانش نیست. تنها منبع دانش تجربه است



شایعات فضای
مجازی زیادند،
بی تحقیق، **باور نکنید!**



امام علی علیه السلام

هرچه را شنیدی (بدون بررسی) برای مردم بازگو مکن
که همین برای دروغگویان تو کافی است

اگر می‌توانستیم خودم مسئولیت این شورا (شورای عالی فضای
مجازی) را به عهده می‌گرفتم





IUVMPixel.COM

the guardian

Like birds of prey circling high in the desert sky, Iran's many foes and rivals are watching the street protests in Tehran and other cities with beady-eyed anticipation.

هر دو تا شلوار را یک دشمن پاره کرده!



تاریخچه

www.bigdata-ir.com

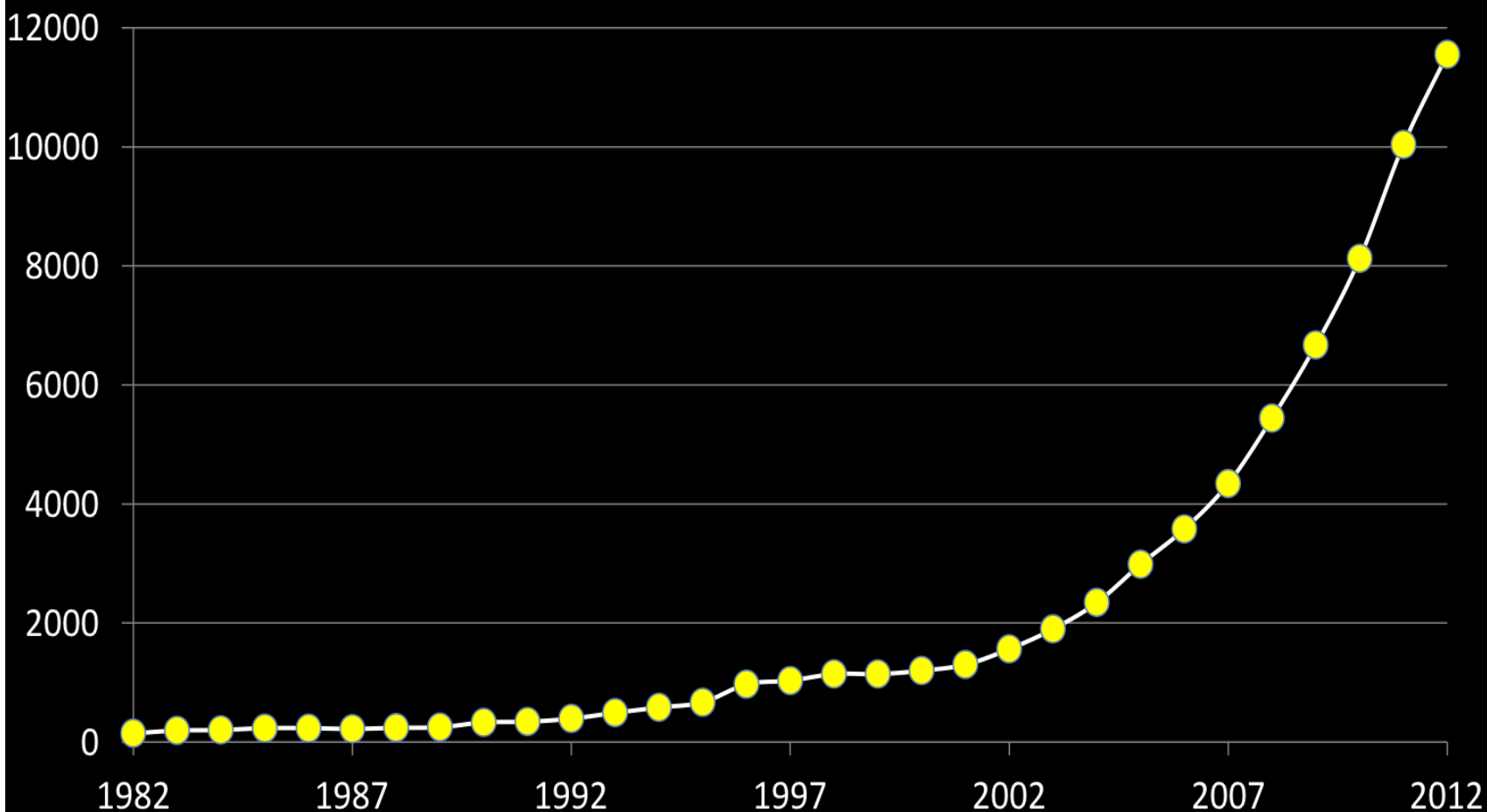
Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

رشد توجه دنیا به حوزه شبکه‌های اجتماعی

تعداد مقالات در حوزه شبکه‌های اجتماعی



◎ ۸۰ سال از عمر شبکه های اجتماعی میگذرد

مقدمه

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

علاقه‌مندی کشورهای مختلف در انتخاب شبکه‌های اجتماعی

The Most Time Consuming Social Networks

Social networks with the most minutes spent on average per user per month in the U.S.*

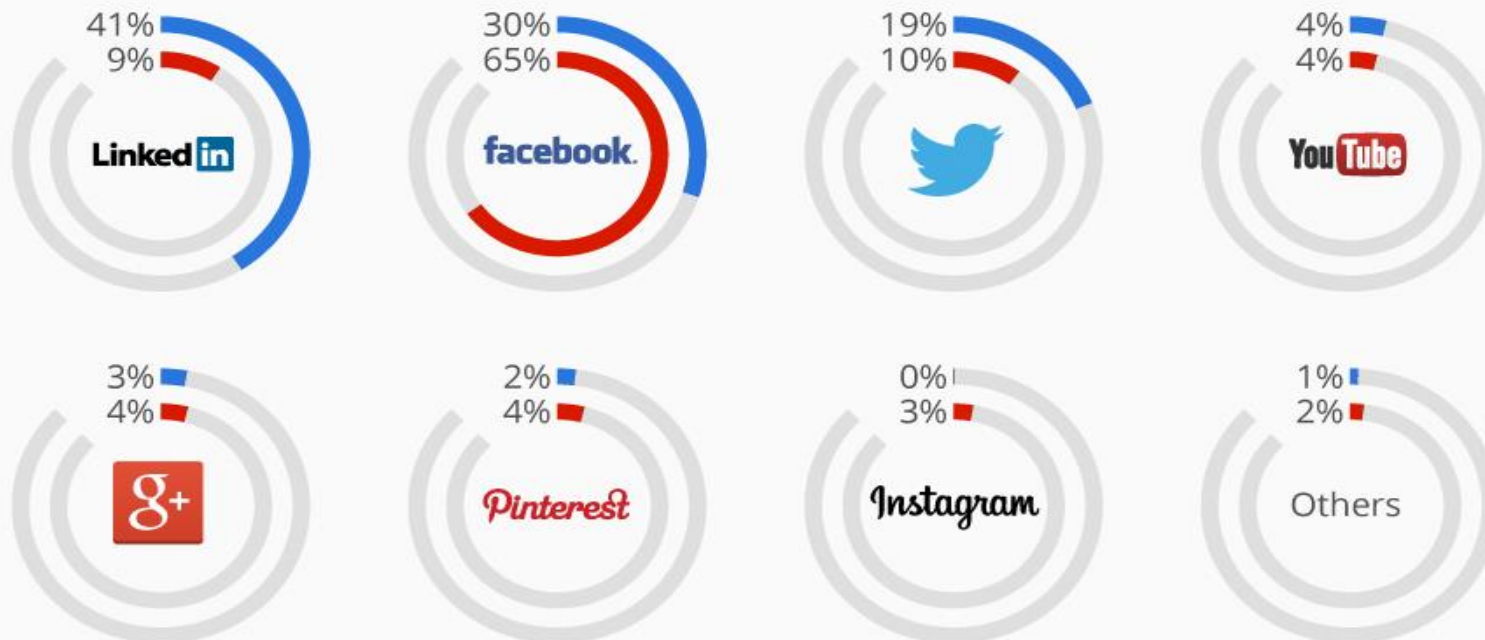


فعالیت تجاری بر روی شبکه ها

B2B Marketers Choose LinkedIn Over Facebook

% of respondents naming platform the single most important for their business

● B2B Marketers ● B2C Marketers

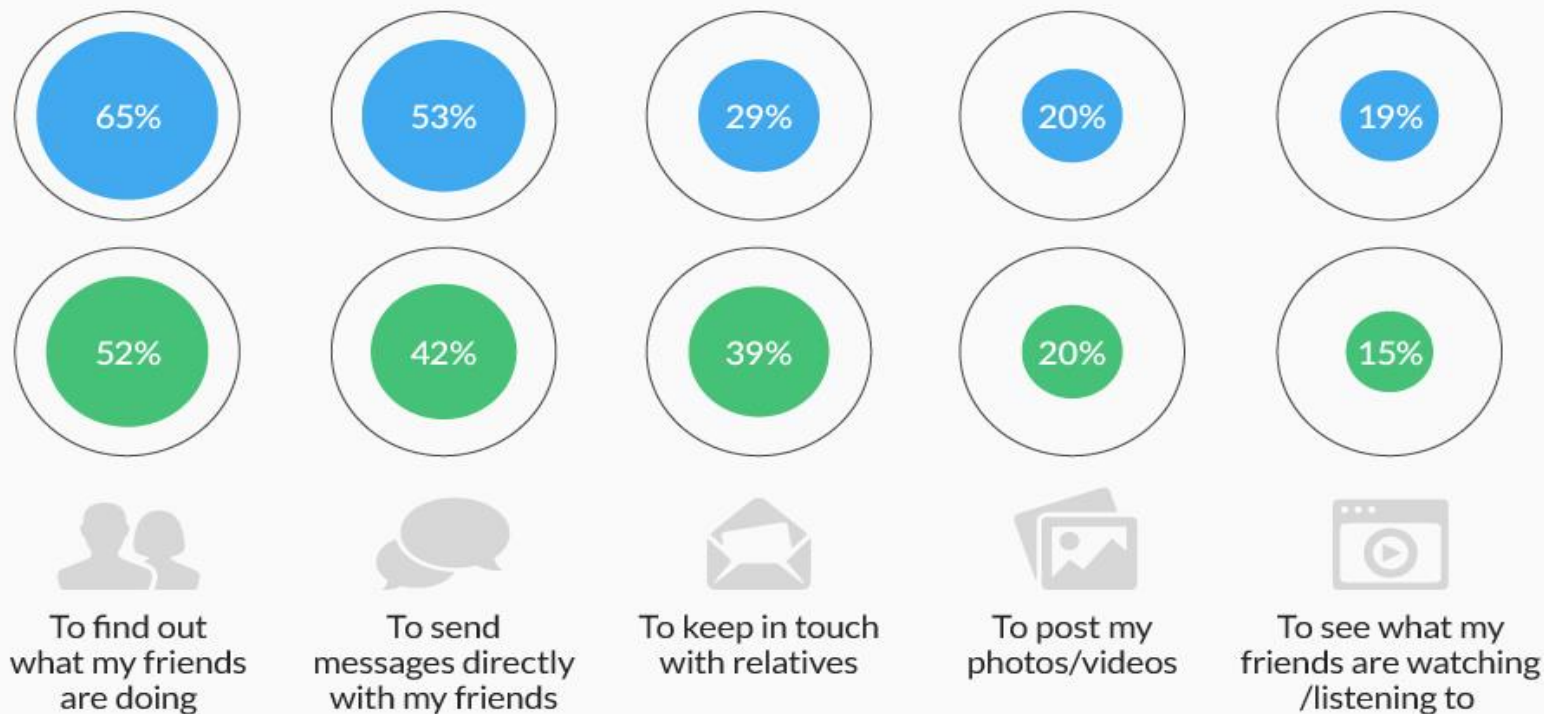


برخی از علت حضور در شبکه ها

Why do people use social media?

The biggest motivations behind social media use in the UK in 2015

16-24s 25-34s



@StatistaCharts

Source: Deloitte

i100

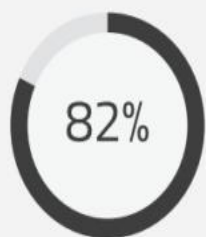
from The INDEPENDENT

statista

بازدید کننده ها در برابر مشارکت کننده ها

ACTIVE USERS VS VISITORS

% who visited the site last month



YouTube



Facebook



Twitter



Google+

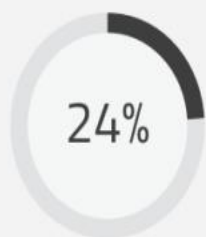


Linkedin



Badoo

% who used or contributed to the site in the last month



YouTube



Facebook



Twitter



Google+

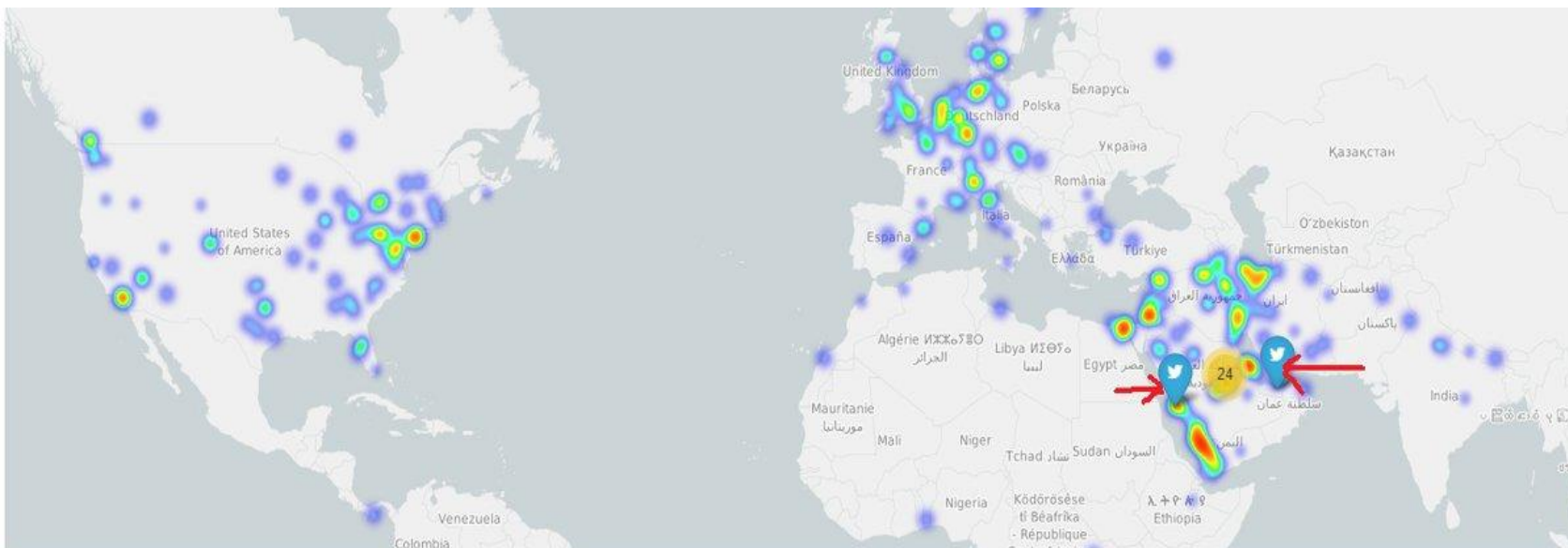


Linkedin



Badoo

نقش جریان های خارج کشور در حوادث اخیر (۹۶)



Conversation (%)

166%	Iran
58%	Afghanistan
17%	Yemen
15%	Saudi Arabia
15%	Sweden
14%	Israel
11%	Syria
8%	Germany

Country

22%	Iran
21%	Saudi Arabia
10%	United Kingdom
7%	France
4%	Canada
4%	Germany
3%	Turkey
2%	United Arab Emirates

City

14%	Tehran, IR
9%	London, GB
9%	Riyadh, SA
2%	Istanbul, TR
2%	Toronto, CA
2%	Paris, FR
2%	Vancouver, CA
1%	Mecca, SA

حوادث اخیر

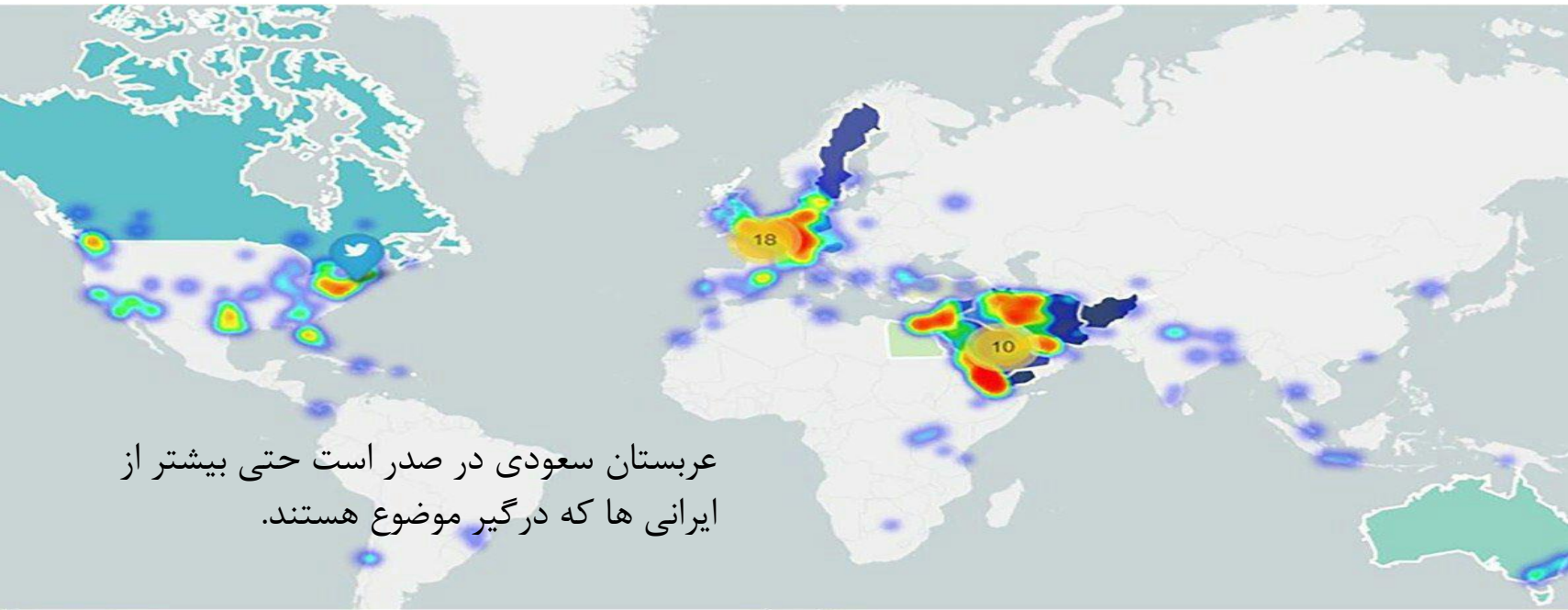
○ از مجموع توییت‌ها با هشتگ تظاهرات سراسری فقط ۲۶٪ توییت‌ها از داخل ایران و ۷۴٪ توییت‌ها خارج از ایران بوده است. که ۲۷٪ آن از عربستان بوده!!!



بیش از ۷۰٪ توییت‌های مربوط به #تظاهرات سراسری خارج از ایران توییت شده‌اند
بیش از ۷۰٪ توییت‌های مربوط به #تظاهرات سراسری خارج از ایران توییت شده‌اند



فراوانی توئیت‌ها در سراسر جهان



عربستان سعودی در صدر است حتی بیشتر از
ایرانی ها که درگیر موضوع هستند.

Country 📍

25%	Saudi Arabia
21%	Iran
7%	United Kingdom
5%	France
4%	Germany
3%	Canada
2%	United Arab Emirates

City 📍

13%	Tehran, IR
10%	Riyadh, SA
6%	London, GB
2%	Mecca, SA
2%	Paris, FR
1%	Toronto, CA
1%	Istanbul, TR

تحليل شبکه های اجتماعی بدون اجرای حتی یک الگوریتم گراف کاوی

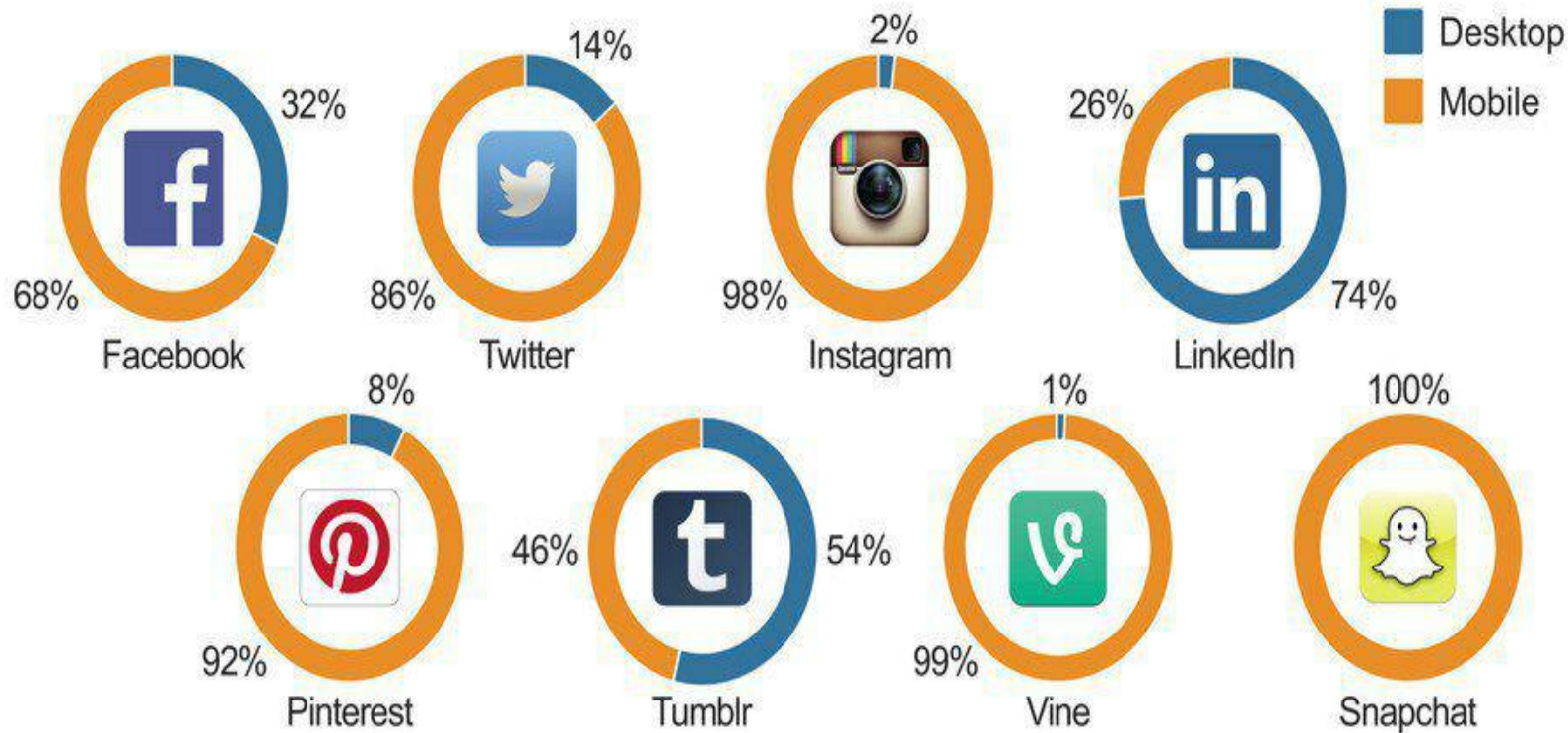
- حساب فارسی سرکرده مجاهدین خلق با ۱۸ هزار دنبال کننده و حساب عربی با ۵۵ هزار دنبال کننده ؟؟؟؟؟!!!!



شبکه های در موبایل و کامپیوتر

Social Network Activity: Mobile vs. Desktop

% of time spent on social networks in the United States, by platform*

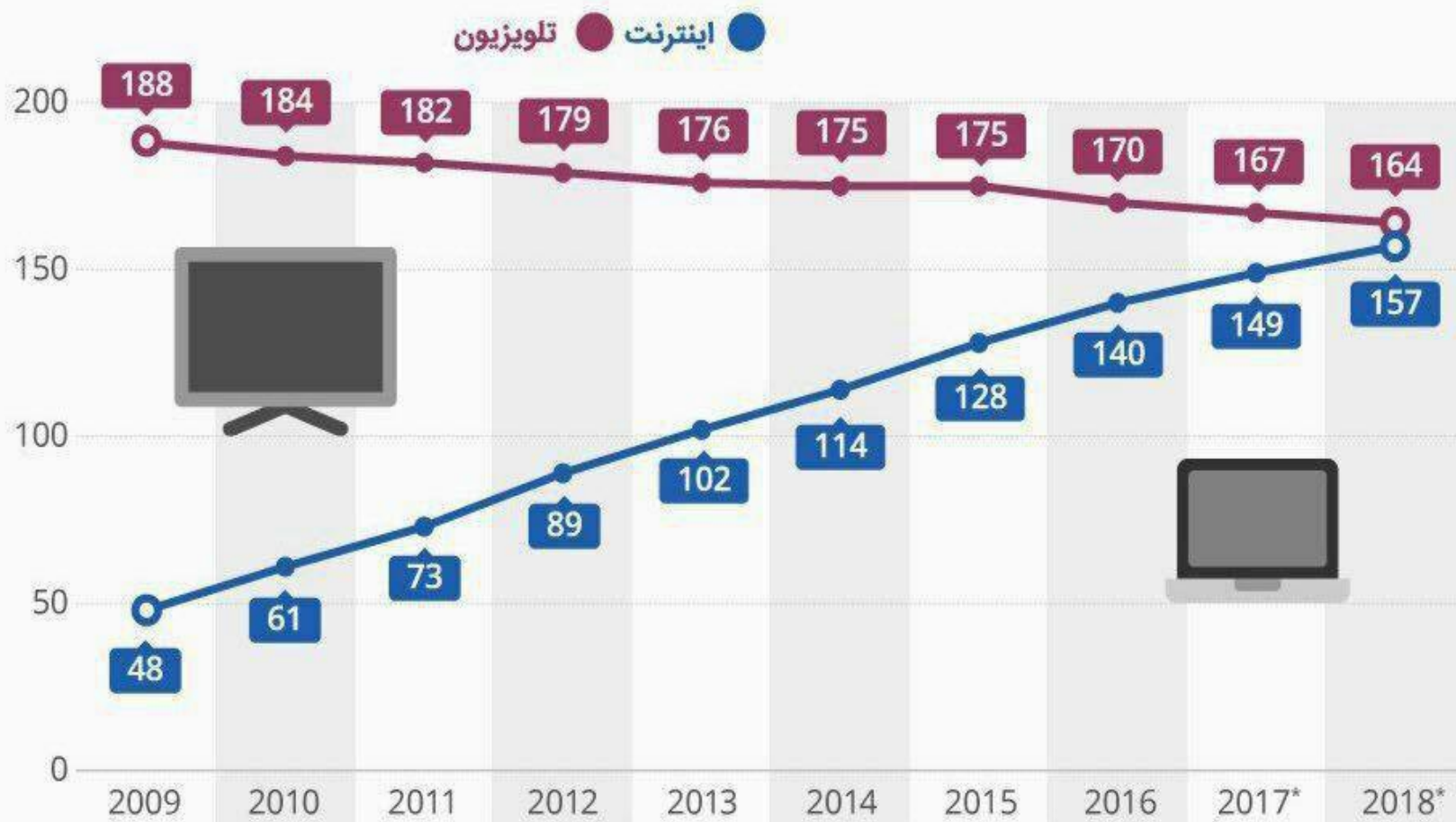


THE WALL STREET JOURNAL. * December 2013, Age 18+

Source: comScore **statista**

و پایان محبوبیت تلویزیون

آیا سلطنت تلویزیون رو به پایان است؟
متوسط استفاده روزانه از **تلویزیون** و **اینترنت** برای هر نفر در سراسر دنیا



Graph theory

تئوری گراف

لینک به مطلب

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

فهرست مطالب

تئوری گراف

انواع گراف

- وجودیت اصلی گراف
- انواع ارتباط ها در گراف
- ماتریس مجاورتی گراف
- وزن یال ها
- درجه نودها
- مسیر در یک گراف
- فاصله در گراف

خصوصیات دیگر

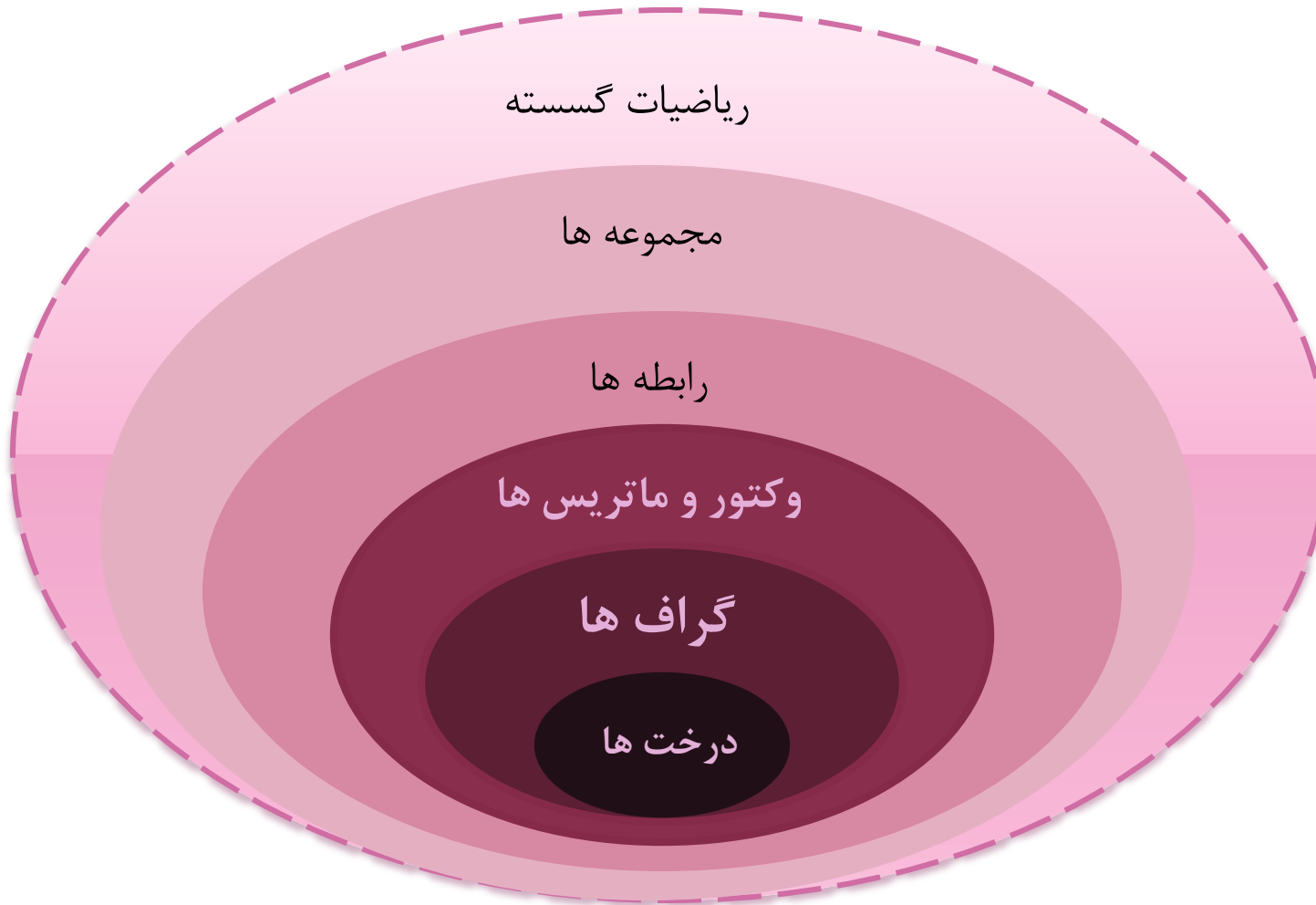
گراف

- گراف جهت دار
- گراف وزن دار
- گراف برچسب دار

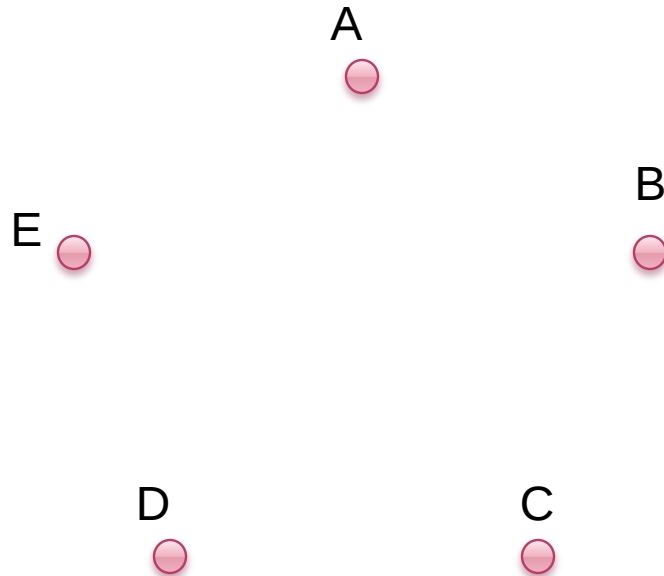
انواع گراف

- گراف کامل
- گراف دو بخشی
- گراف تصادفی
- گراف اسپارس

مروری بر حوزه تحقیق



موجودیت اصلی گراف



ارتباط ها

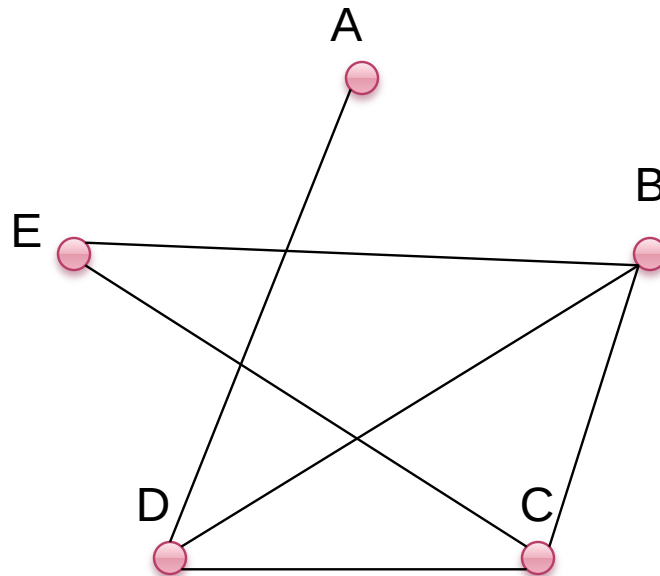
$G(N,E)$

$N=$

تعداد نودها

$E=$

تعداد یال ها



N
 E
 $G(N,E)$

- **Objects:** nodes, vertices
- **Interactions:** links, edges
- **System:** network, graph

گراف جهت دار

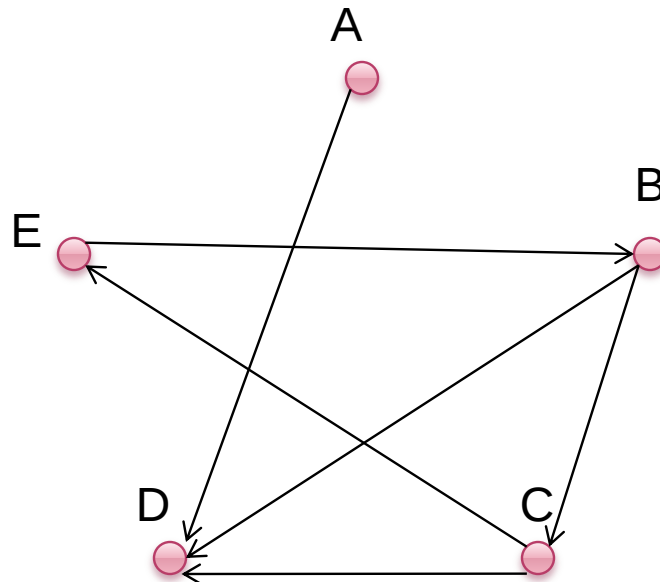
$G(N,E)$

$N=$

تعداد نودها

$E=$

تعداد یال ها



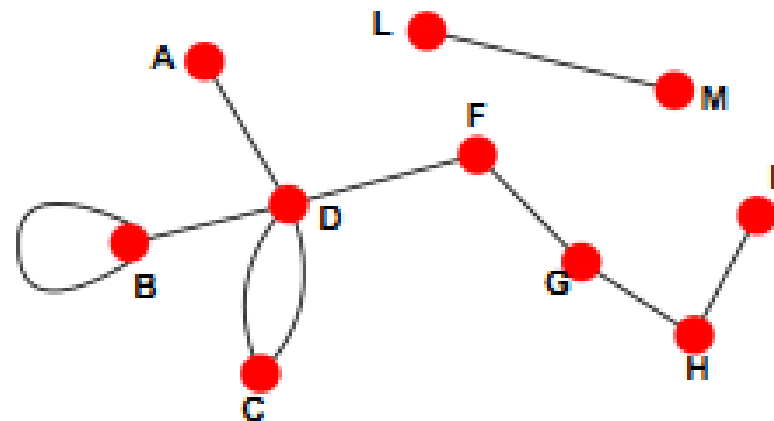
خصوصیات یال ها

- جهت
- وزن
- علامت
- نوع رابطه
- ...

Undirected vs. Directed Networks

Undirected

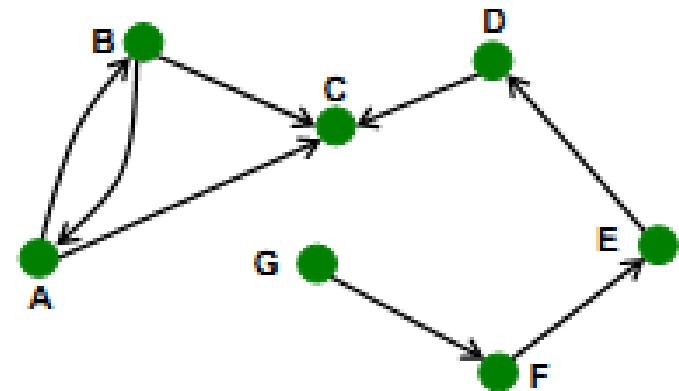
- Links: undirected (symmetrical)



- Examples:
 - Collaborations
 - Friendship on Facebook

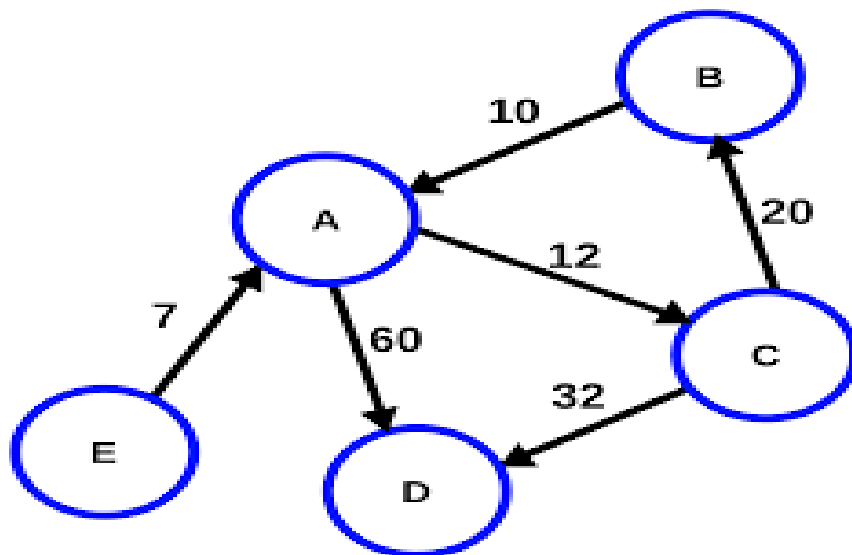
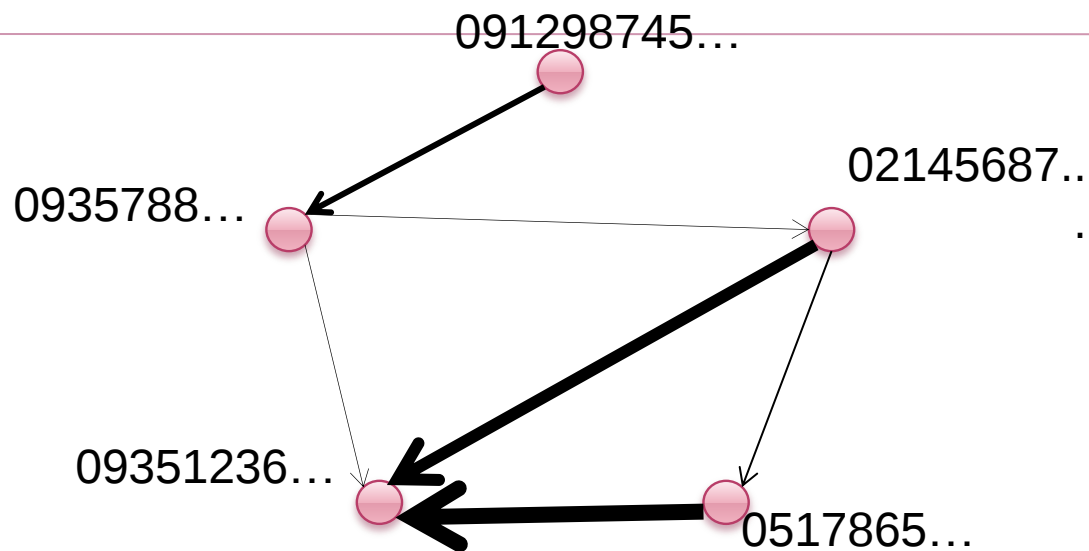
Directed

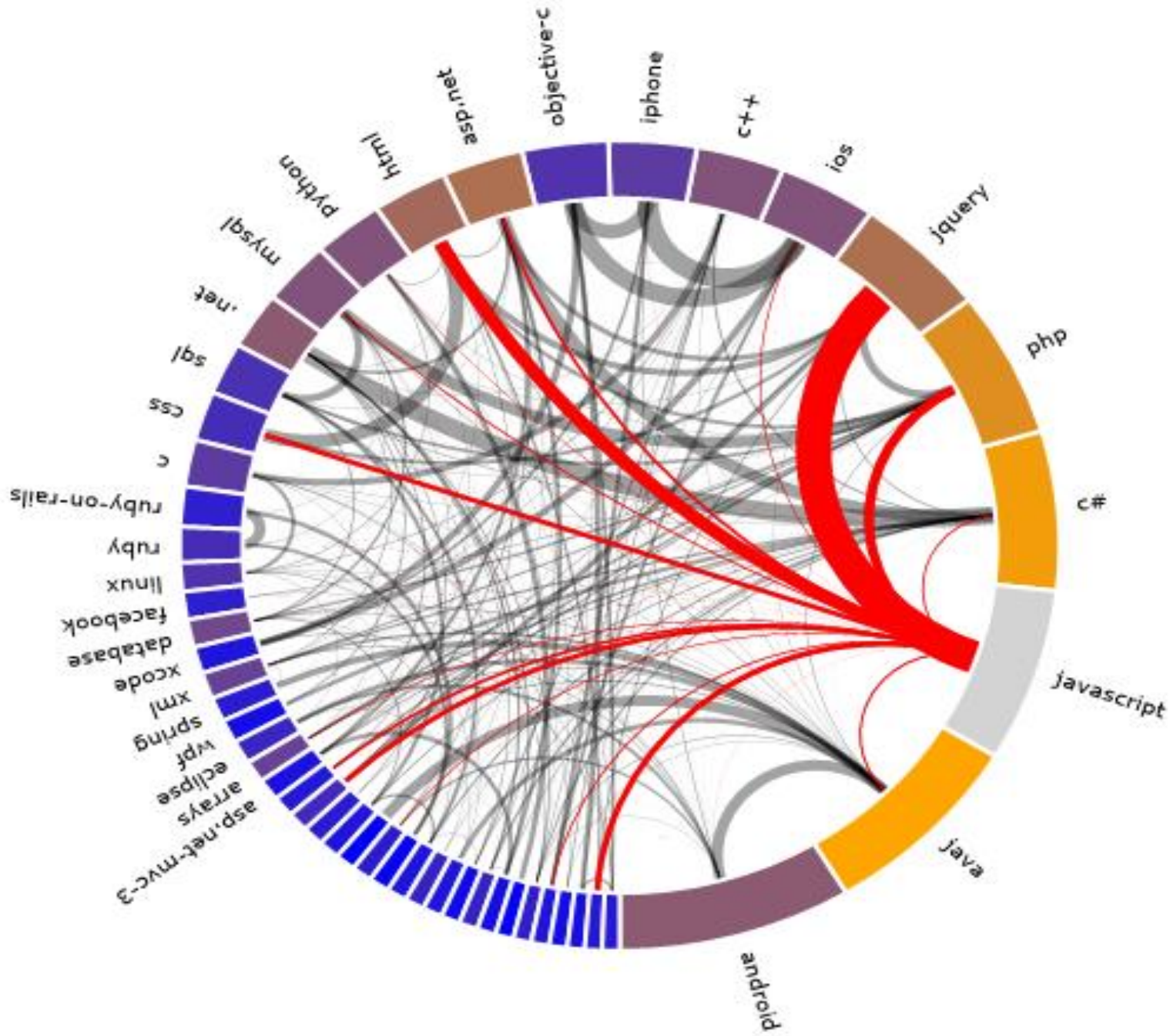
- Links: directed (arcs) Asymmetrical



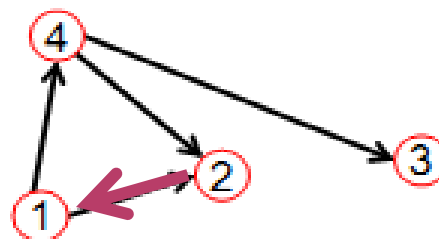
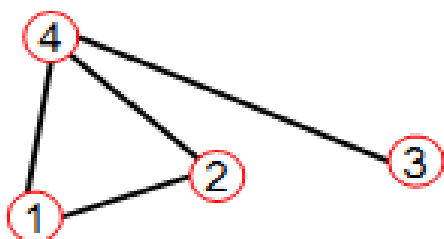
- Examples:
 - Phone calls
 - Following on Twitter

وزن یال ها





Adjacency Matrix



$A_{ij} = 1$ if there is a link from node i to node j

$A_{ij} = 0$ otherwise

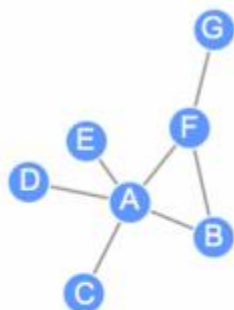
$$A = \begin{pmatrix} 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 \\ 1 & 1 & 1 & 0 \end{pmatrix}$$

$$A = \begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \end{pmatrix}$$

Note that for a directed graph (right) the matrix is not symmetric.

ماتریس مجاور

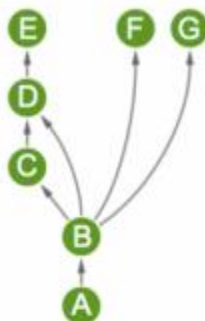
Undirected



	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

Adjacency matrices

Directed



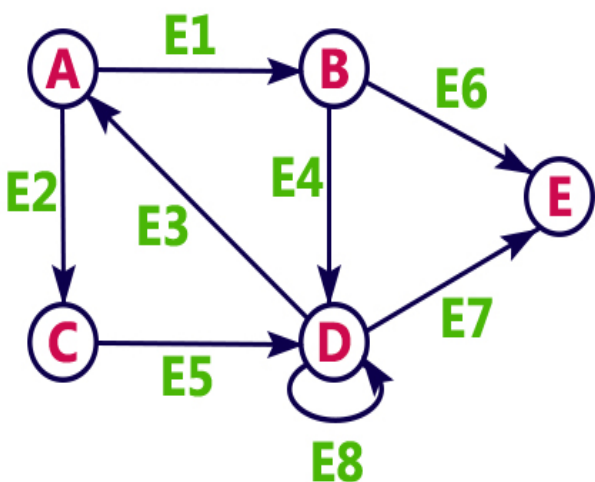
	A	B	C	D	E	F	G	Out-degree
A	1	0	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

Weighted



	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

ماتریس برخورد

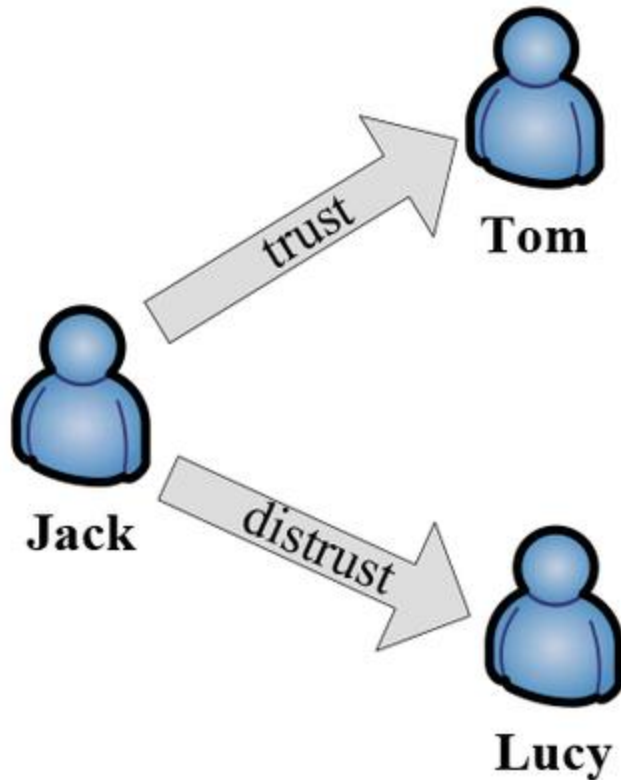


→

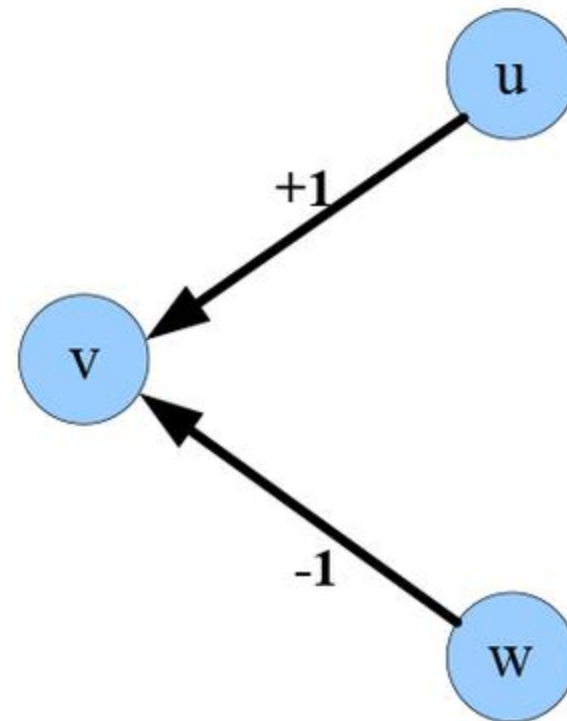
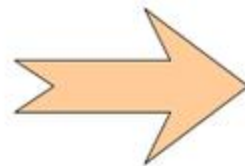
	E1	E2	E3	E4	E5	E6	E7	E8
A	1	1	-1	0	0	0	0	0
B	-1	0	0	1	0	1	0	0
C	0	-1	0	0	1	0	0	0
D	0	0	1	-1	-1	0	1	1
E	0	0	0	0	0	-1	-1	0



SIGNED NETWORK GRAPH



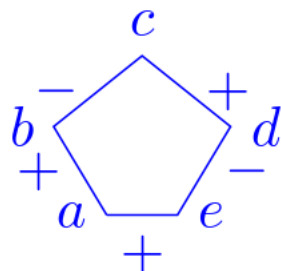
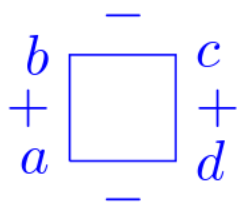
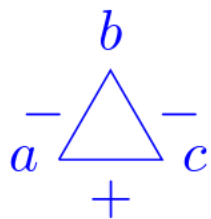
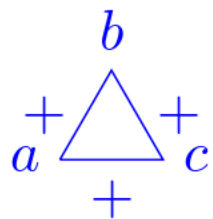
(a) An signed social network



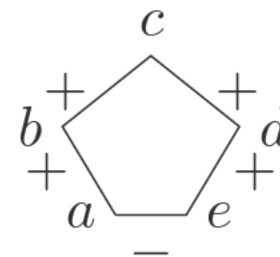
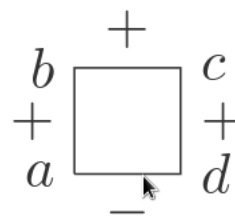
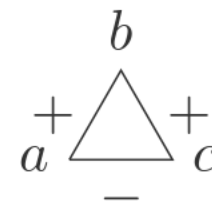
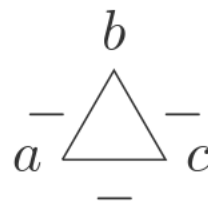
(b) A signed and directed graph

SOCIAL BALANCE THEORY

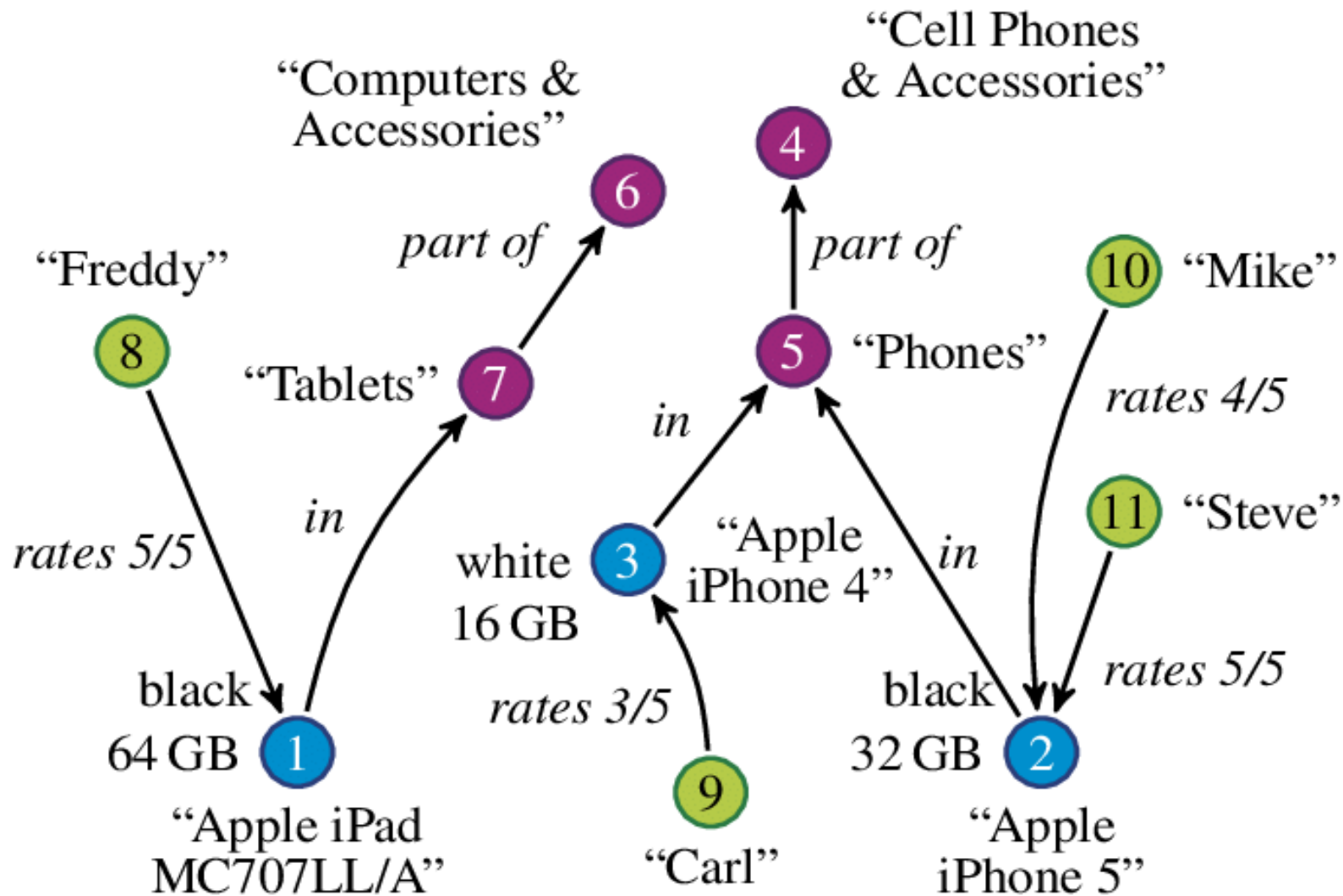
Balanced local patterns
(more preferred)



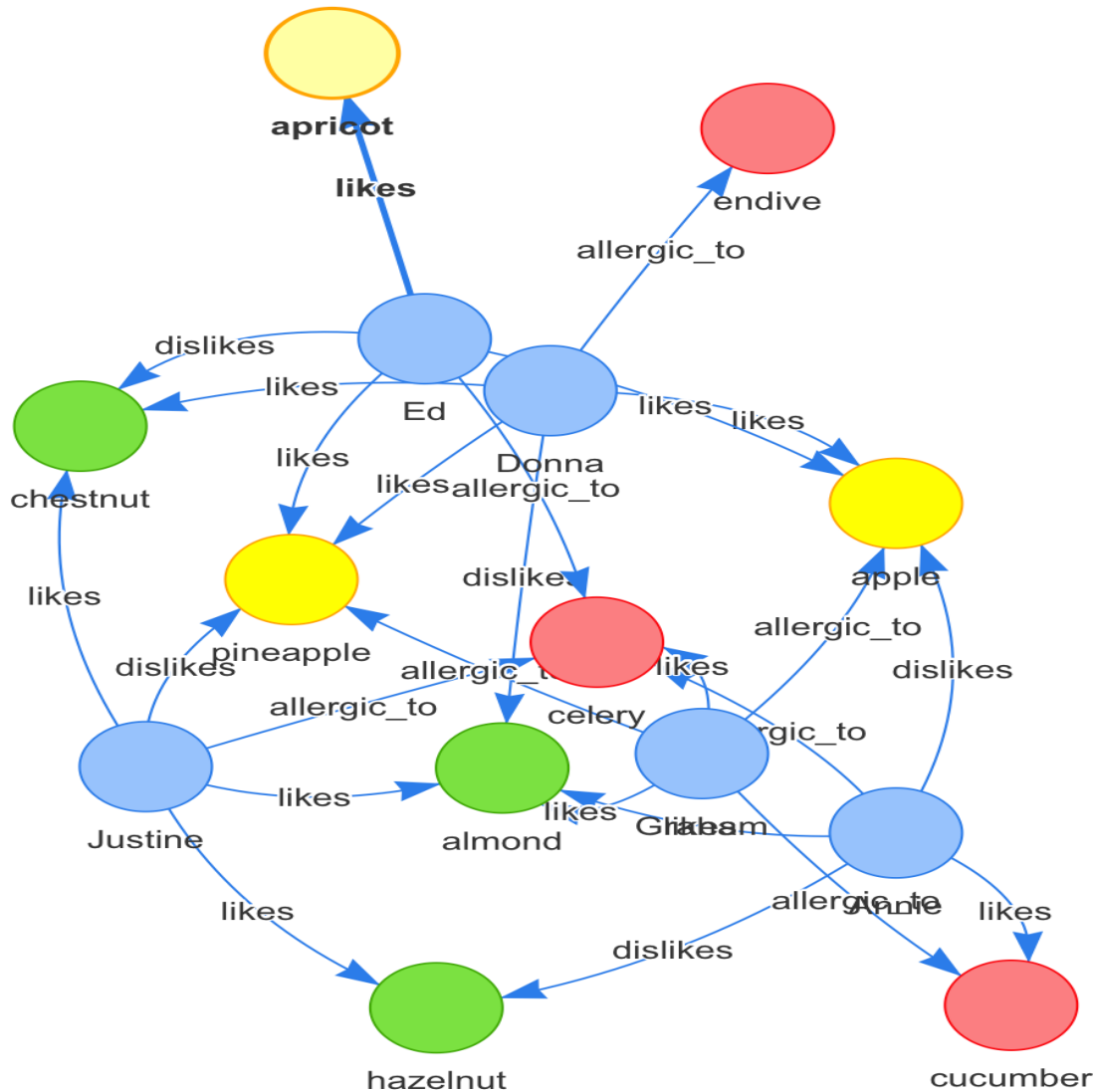
Unbalanced local patterns
(less preferred)



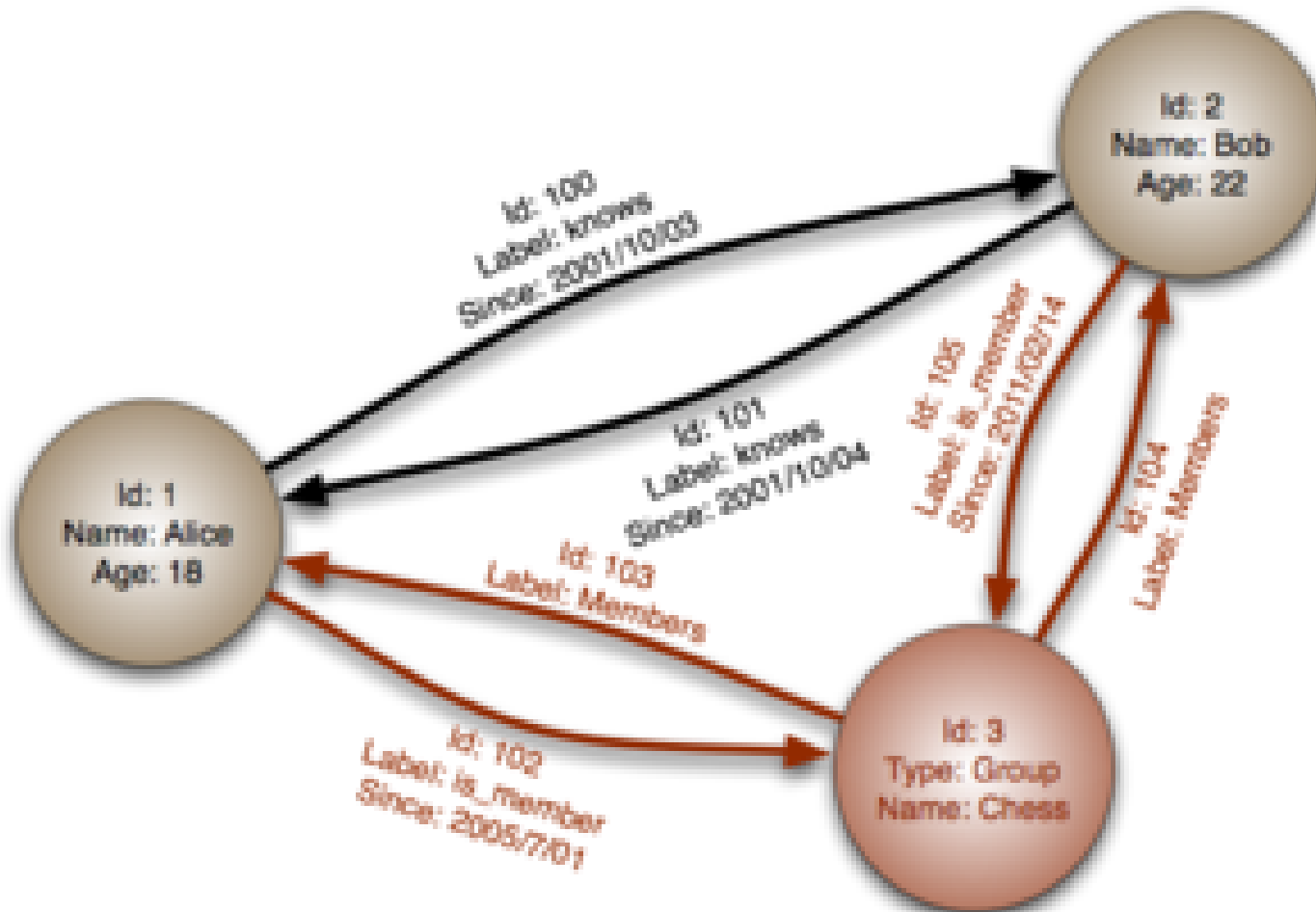
EDGE ATTRIBUTE



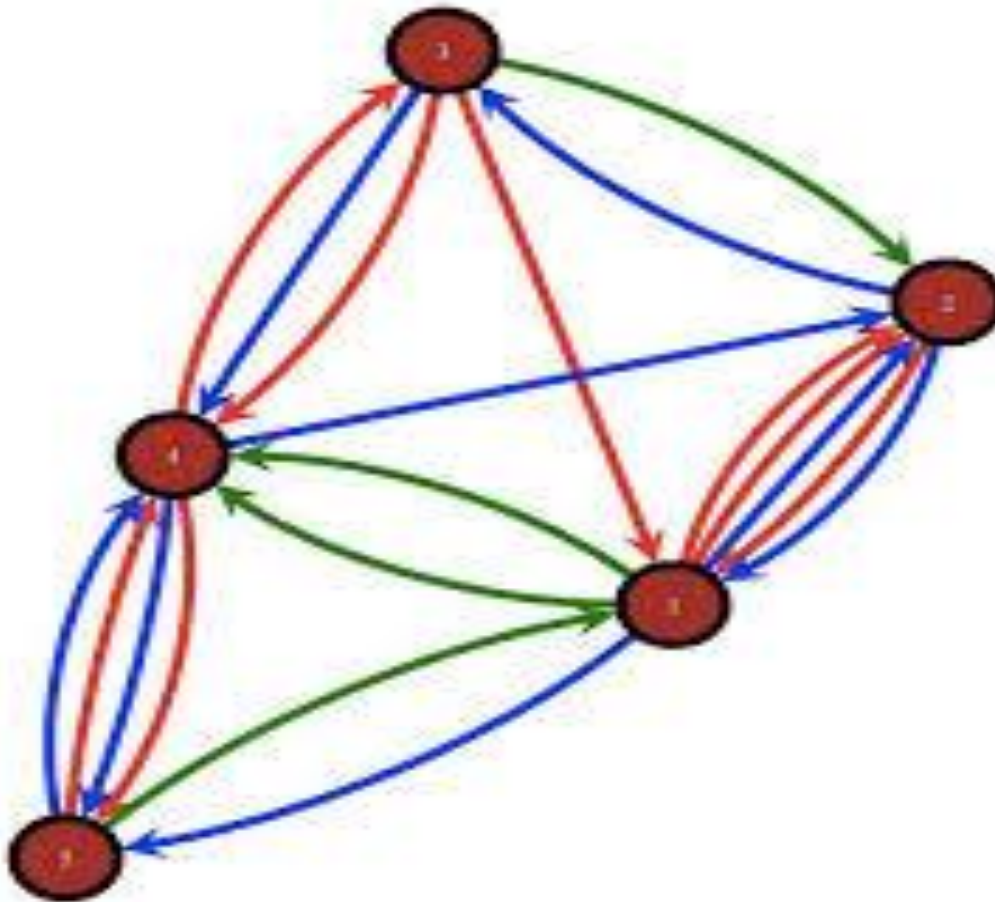
LIKE & DISLIKE EDGE ATTRIBUTE



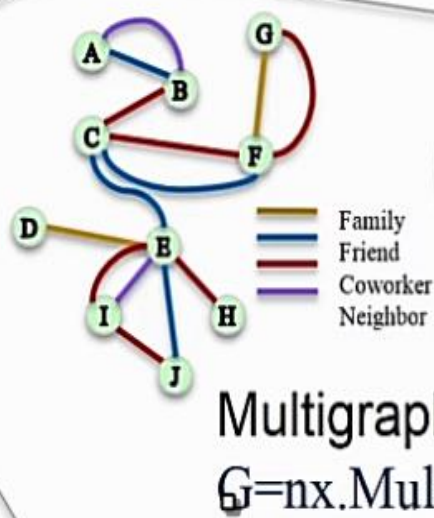
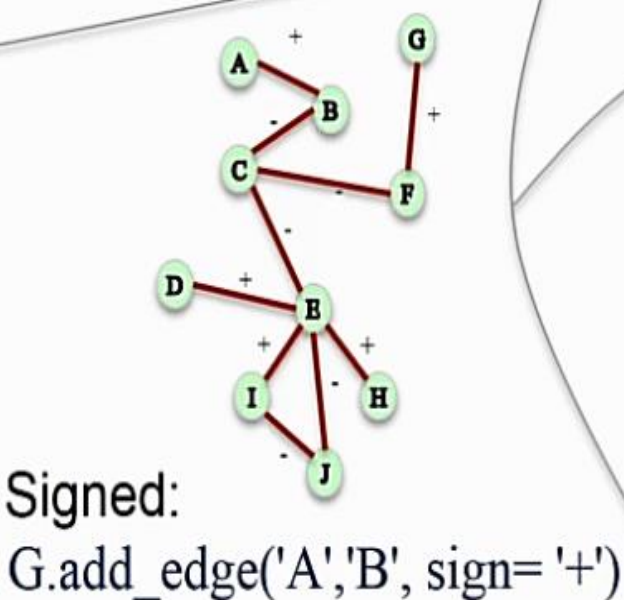
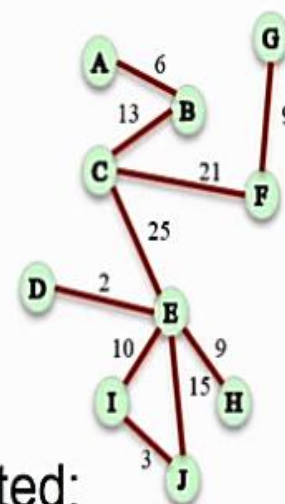
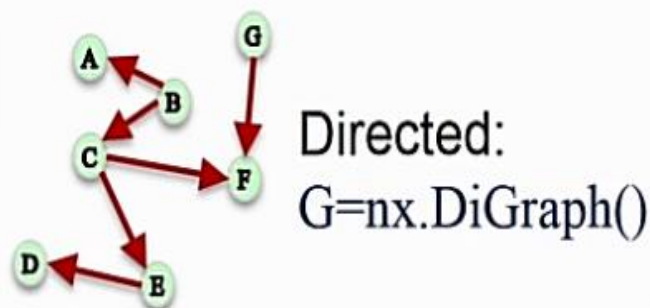
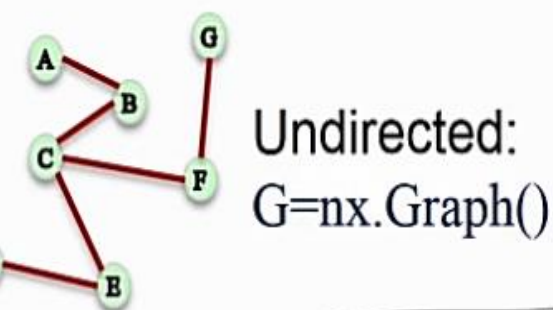
DATE EDGE ATTRIBUTE



MULTI GRAPH

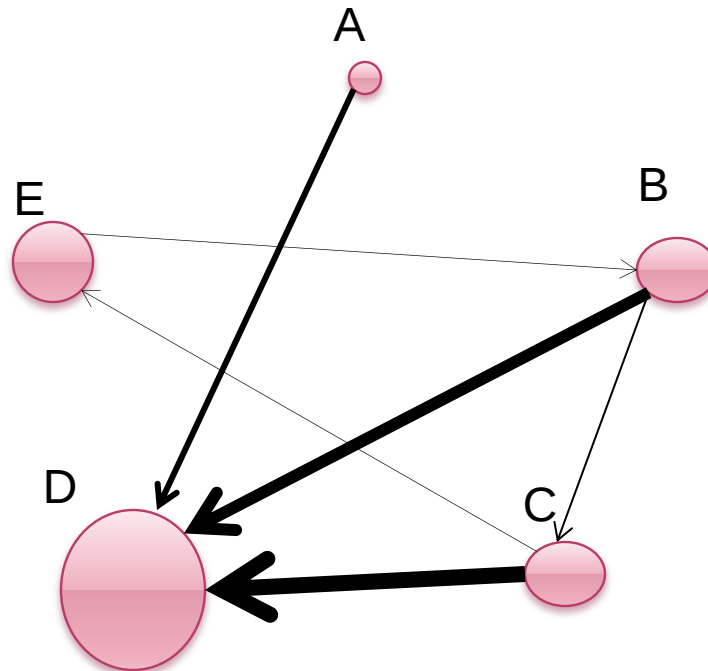


Summary



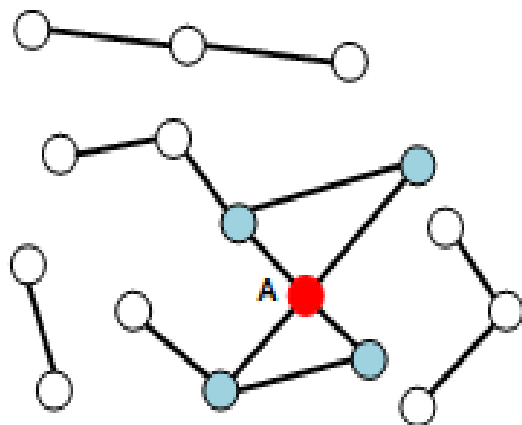
درجه نودها

هر نود میتواند برچسب و مرتبه یا درجه داشته باشد.
که این درجه در مراحل گراف کاوی استخراج خواهد شد.



Node Degrees

Undirected

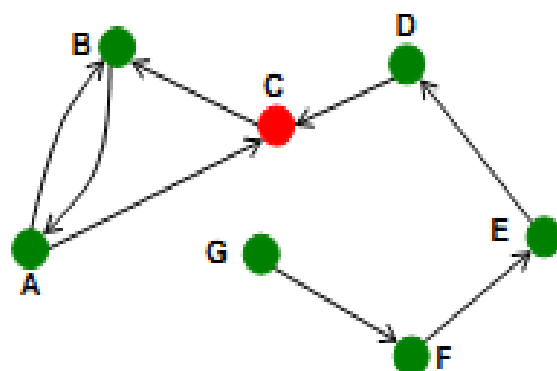


Node degree, k_i : the number of edges adjacent to node i

$$k_A = 4$$

Avg. degree: $\bar{k} = \langle k \rangle = \frac{1}{N} \sum_{i=1}^N k_i = \frac{2E}{N}$

Directed



In directed networks we define an **in-degree** and **out-degree**.
The (total) degree of a node is the sum of in- and out-degrees.

$$k_C^{in} = 2 \quad k_C^{out} = 1 \quad k_C = 3$$

Source: node with $k^{in} = 0$

Sink: node with $k^{out} = 0$

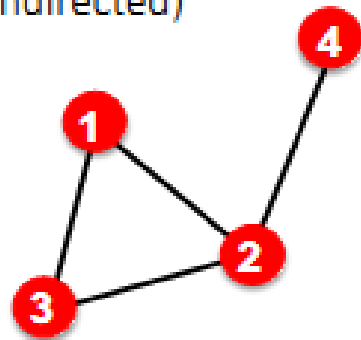
$$\bar{k} = \frac{E}{N}$$

$$\overline{k^{in}} = \overline{k^{out}}$$

More Types of Graphs:

■ Unweighted

(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$A_{ij} = A_{ji}$$

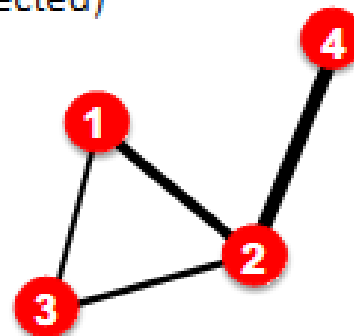
$$E = \frac{1}{2} \sum_{i,j=1}^N A_{ij} \quad \bar{k} = \frac{2E}{N}$$

Examples: Friendship, Sex

ارائه دهنده: علیرضا حبیبی

■ Weighted

(undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 0.5 & 0 \\ 2 & 0 & 1 & 4 \\ 0.5 & 1 & 0 & 0 \\ 0 & 4 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$A_{ij} = A_{ji}$$

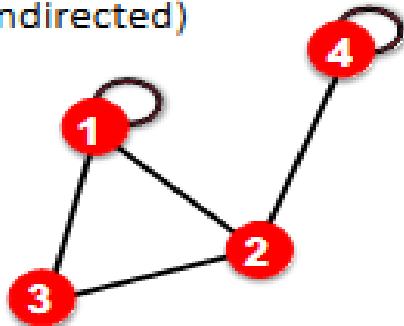
$$E = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \bar{k} = \frac{2E}{N}$$

Examples: Collaboration, Internet, Roads

دوره آموزشی شبکه های اجتماعی

More Types of Graphs:

Self-edges (self-loops) (undirected)



$$A_{ij} = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}$$

$$A_{ii} \neq 0$$

$$A_{ij} = A_{ji}$$

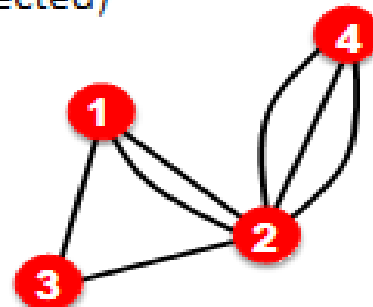
?

$$E = \frac{1}{2} \sum_{i,j=1, i \neq j}^N A_{ij} + \sum_{i=1}^N A_{ii}$$

Examples: Proteins

ارائه دهنده: علیرضا حبیبی

Multigraph (undirected)



$$A_{ij} = \begin{pmatrix} 0 & 2 & 1 & 0 \\ 2 & 0 & 1 & 3 \\ 1 & 1 & 0 & 0 \\ 0 & 3 & 0 & 0 \end{pmatrix}$$

$$A_{ii} = 0$$

$$A_{ij} = A_{ji}$$

$$E = \frac{1}{2} \sum_{i,j=1}^N \text{nonzero}(A_{ij}) \quad \bar{k} = \frac{2E}{N}$$

Examples: Communication, Collaboration

دوره آموزشی شبکه های اجتماعی

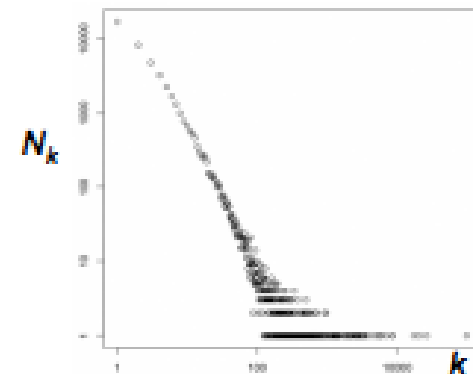
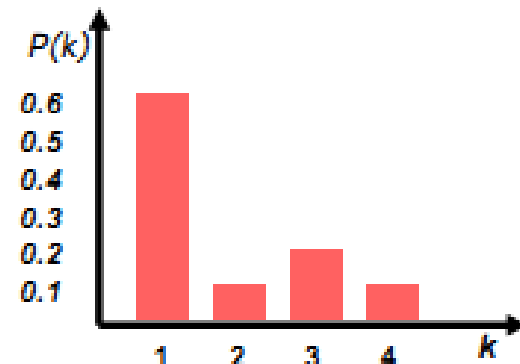
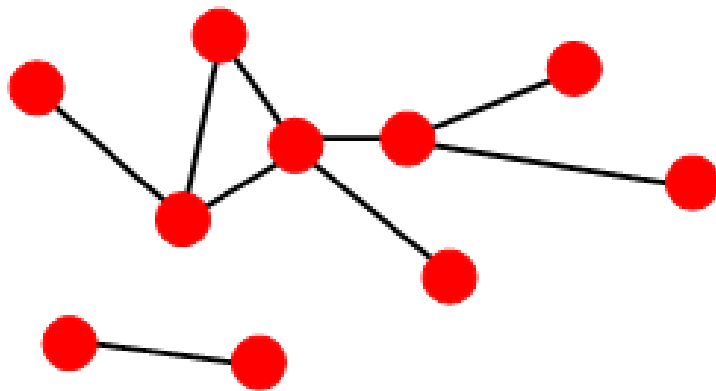
Degree Distribution

- **Degree distribution $P(k)$:** Probability that a randomly chosen node has degree k

$N_k = \#$ nodes with degree k

- **Normalized histogram:**

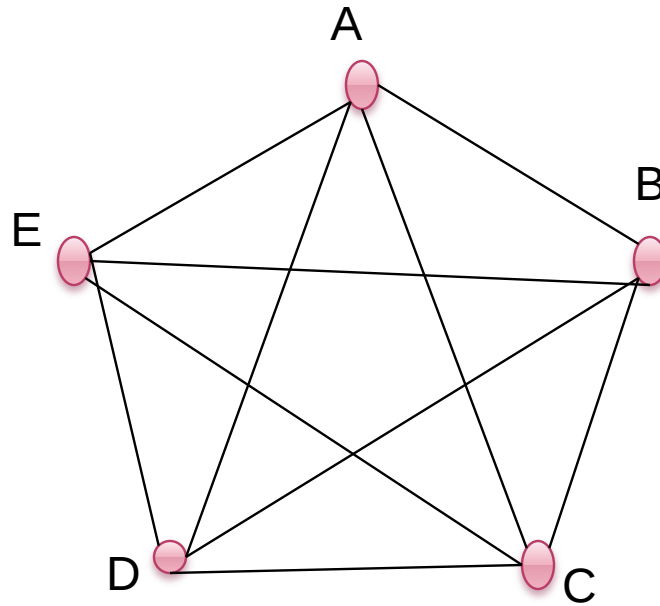
$$P(k) = N_k / N \rightarrow \text{plot}$$



گراف کامل

$$K = n - 1$$

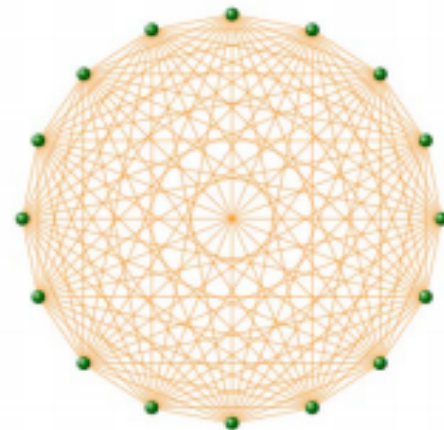
درجه برای همه نودها:



Complete Graph

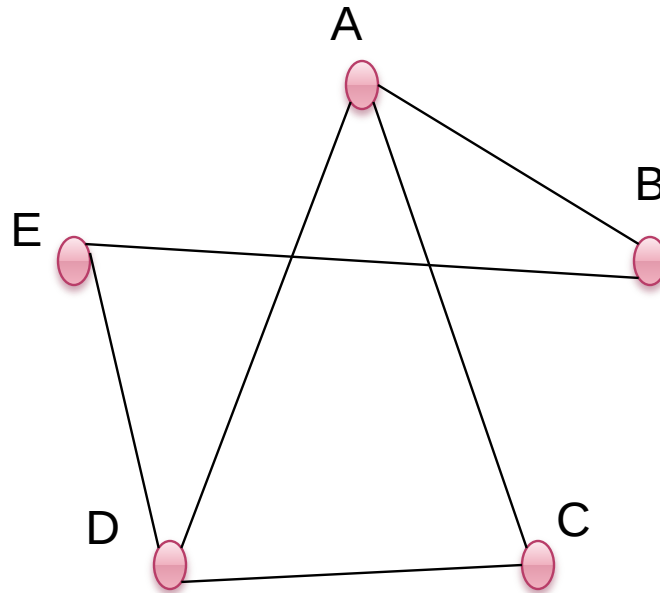
The **maximum number of edges** in an undirected graph on N nodes is

$$E_{\max} = \binom{N}{2} = \frac{N(N-1)}{2}$$



A graph with the number of edges $E = E_{\max}$ is a **complete graph**, and its average degree is $N-1$

گراف اسپارس



Networks are Sparse Graphs

Most real-world networks are **sparse**

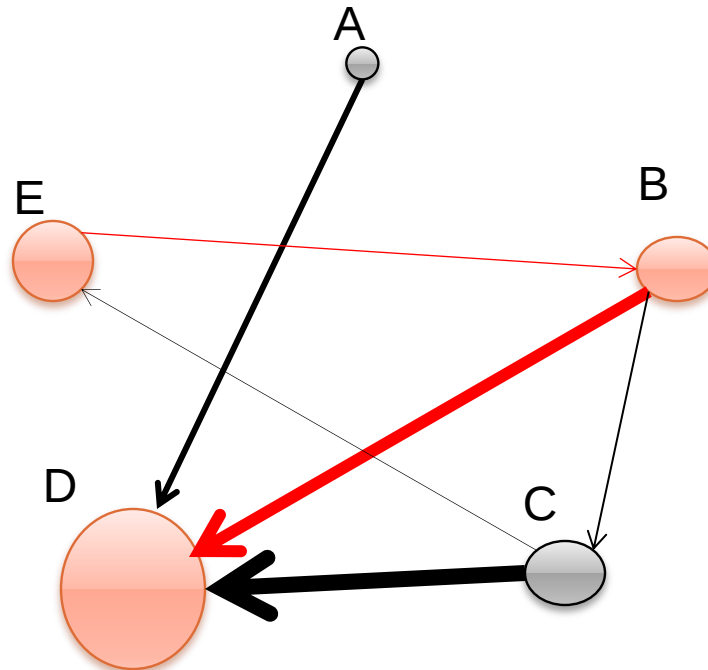
$$E \ll E_{\max} \text{ (or } \bar{k} \ll N-1)$$

WWW (Stanford-Berkeley):	$N=319,717$	$\langle k \rangle=9.65$
Social networks (LinkedIn):	$N=6,946,668$	$\langle k \rangle=8.87$
Communication (MSN IM):	$N=242,720,596$	$\langle k \rangle=11.1$
Coauthorships (DBLP):	$N=317,080$	$\langle k \rangle=6.62$
Internet (AS-Skitter):	$N=1,719,037$	$\langle k \rangle=14.91$
Roads (California):	$N=1,957,027$	$\langle k \rangle=2.82$
Protein (<i>S. Cerevisiae</i>):	$N=1,870$	$\langle k \rangle=2.39$

(Source: Leskovec et al., *Internet Mathematics*, 2009)

Consequence: Adjacency matrix is filled with zeros!

(**Density** (E/N^2): WWW= 1.51×10^{-5} , MSN IM = 2.27×10^{-8})

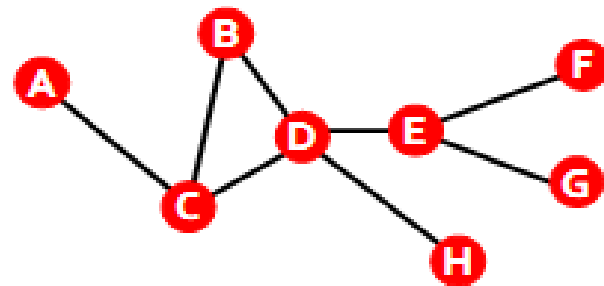


Paths in a Graph

- A **path** is a sequence of nodes in which each node is linked to the next one

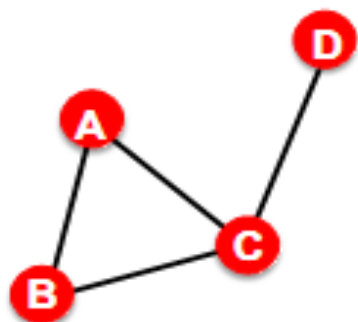
$$P_n = \{i_0, i_1, i_2, \dots, i_n\} \quad P_n = \{(i_0, i_1), (i_1, i_2), (i_2, i_3), \dots, (i_{n-1}, i_n)\}$$

- Path can intersect itself and pass through the same edge multiple times
 - E.g.: ACBDCDEG
 - In a directed graph a path can only follow the direction of the "arrow"



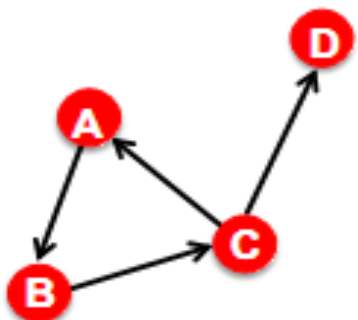
Distance in a Graph

فاصله یعنی کوتاه ترین مسیر



$$h_{B,D} = 2$$

- **Distance (shortest path, geodesic)** between a pair of nodes is defined as the number of edges along the shortest path connecting the nodes
 - *If the two nodes are disconnected, the distance is usually defined as infinite

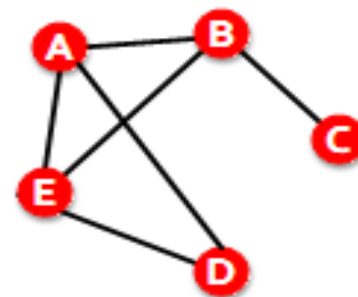
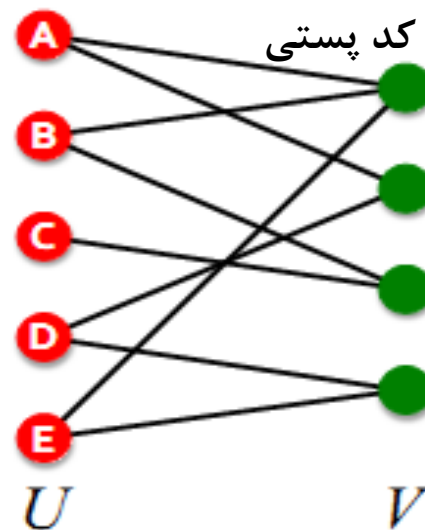


$$h_{B,C} = 1, h_{C,B} = 2$$

- In **directed graphs** paths need to follow the direction of the arrows
 - Consequence: Distance is not symmetric: $h_{A,C} \neq h_{C,A}$

Bipartite Graph

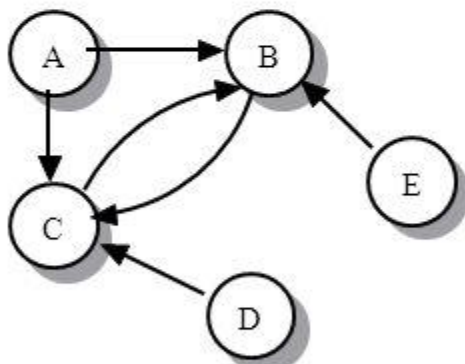
- **Bipartite graph** is a graph whose nodes can be divided into two disjoint sets U and V such that every link connects a node in U to one in V ; that is, U and V are independent sets.
- **Examples:**
 - Authors-to-papers (they authored)
 - Actors-to-Movies (they appeared in)
 - Users-to-Movies (they rated)
- **“Folded” networks:**
 - Author collaboration networks
 - Movie co-rating networks



Strength of Connectedness (digraphs only)

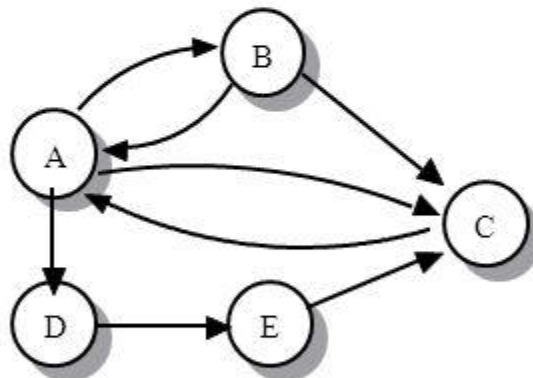


- Strongly connected if there is a path from each vertex to every other vertex.



Not Strongly or Weakly Connected
(No path E to D or D to E)

(a)



Strongly Connected

(b)

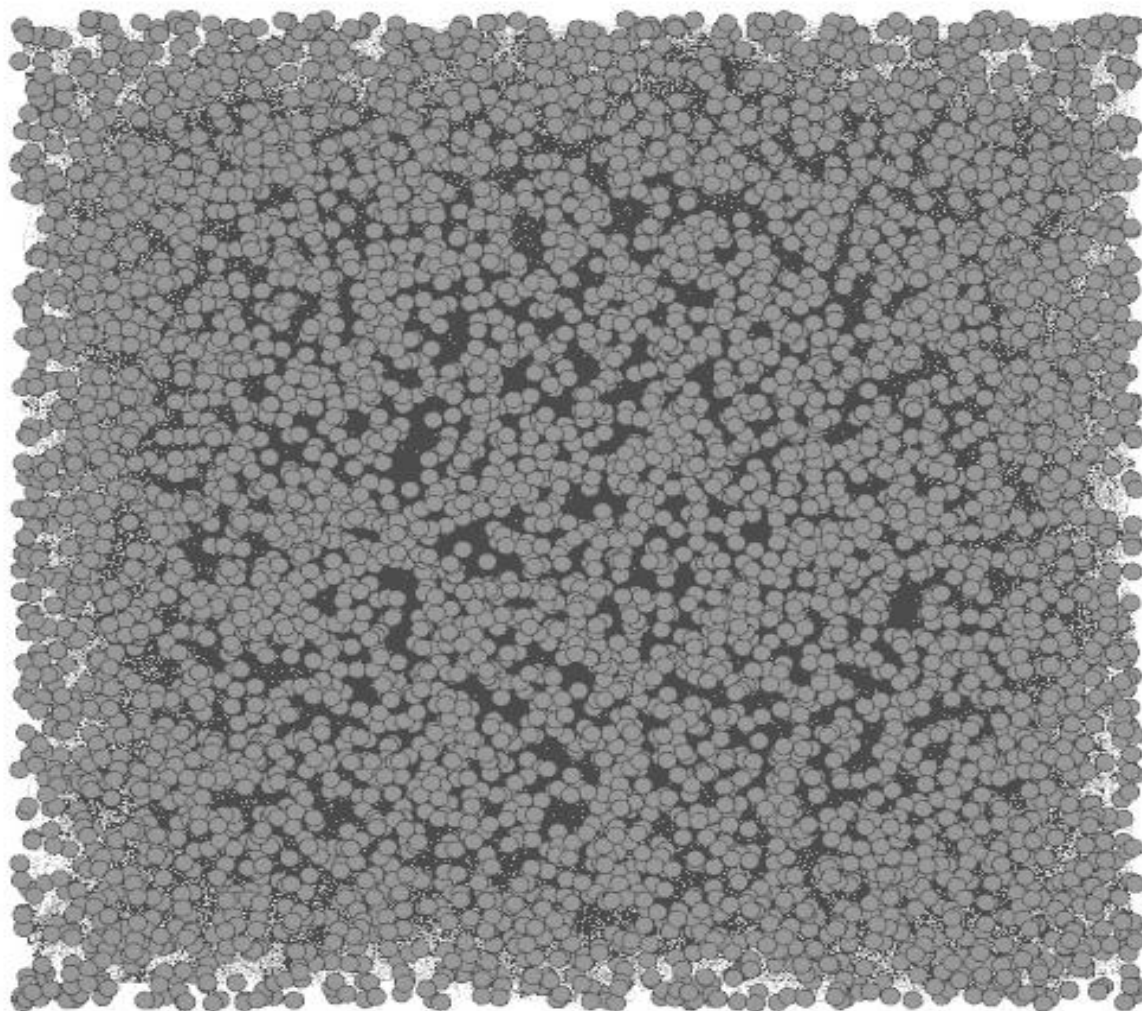
نمونه گراف

www.bigdata-ir.com

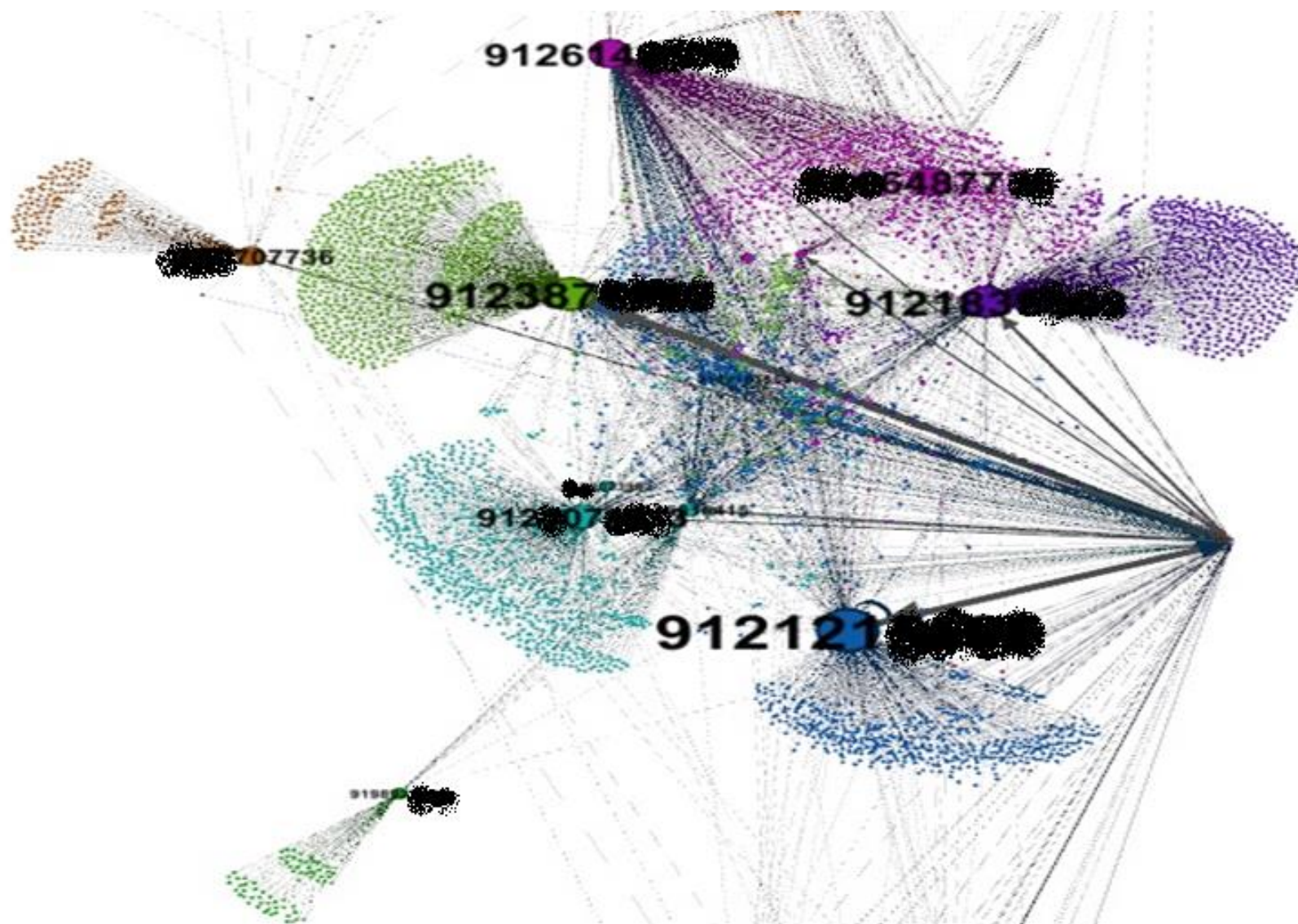
Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:
علیرضا حبیبی

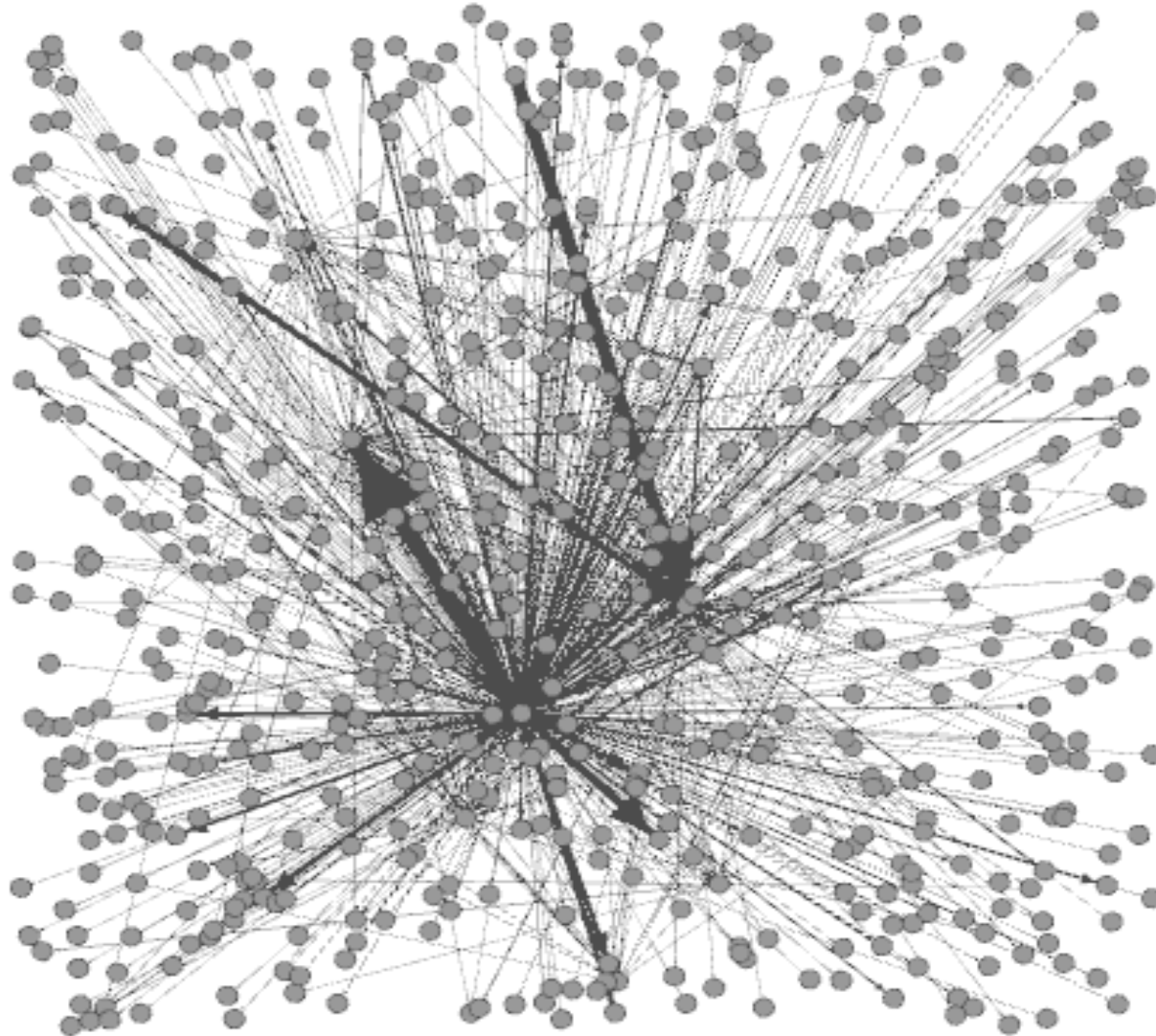
گراف ارتباطات تلفنی (قبل از تحلیل)



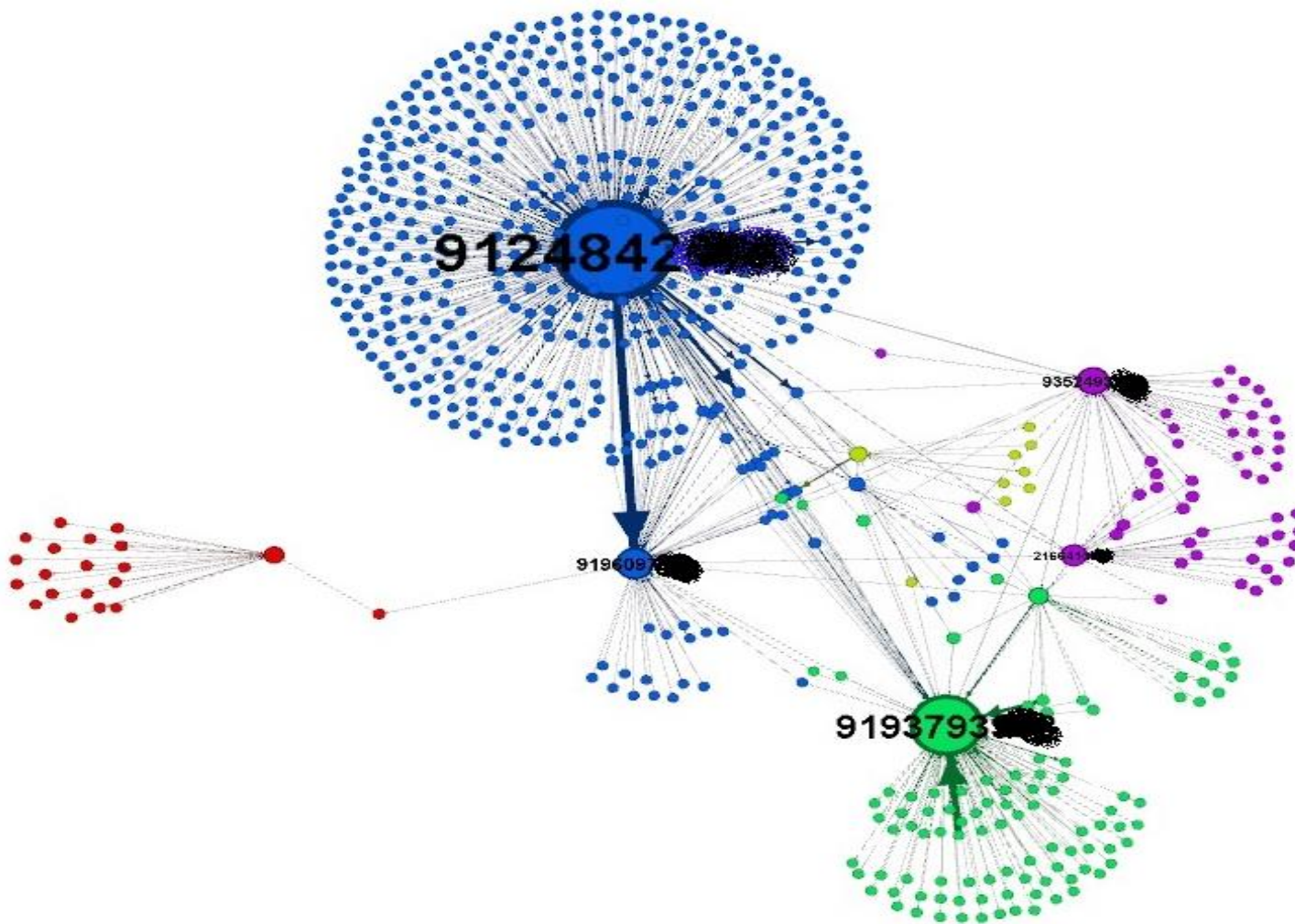
گراف ارتباطات تلفنی (بعد از تحلیل)

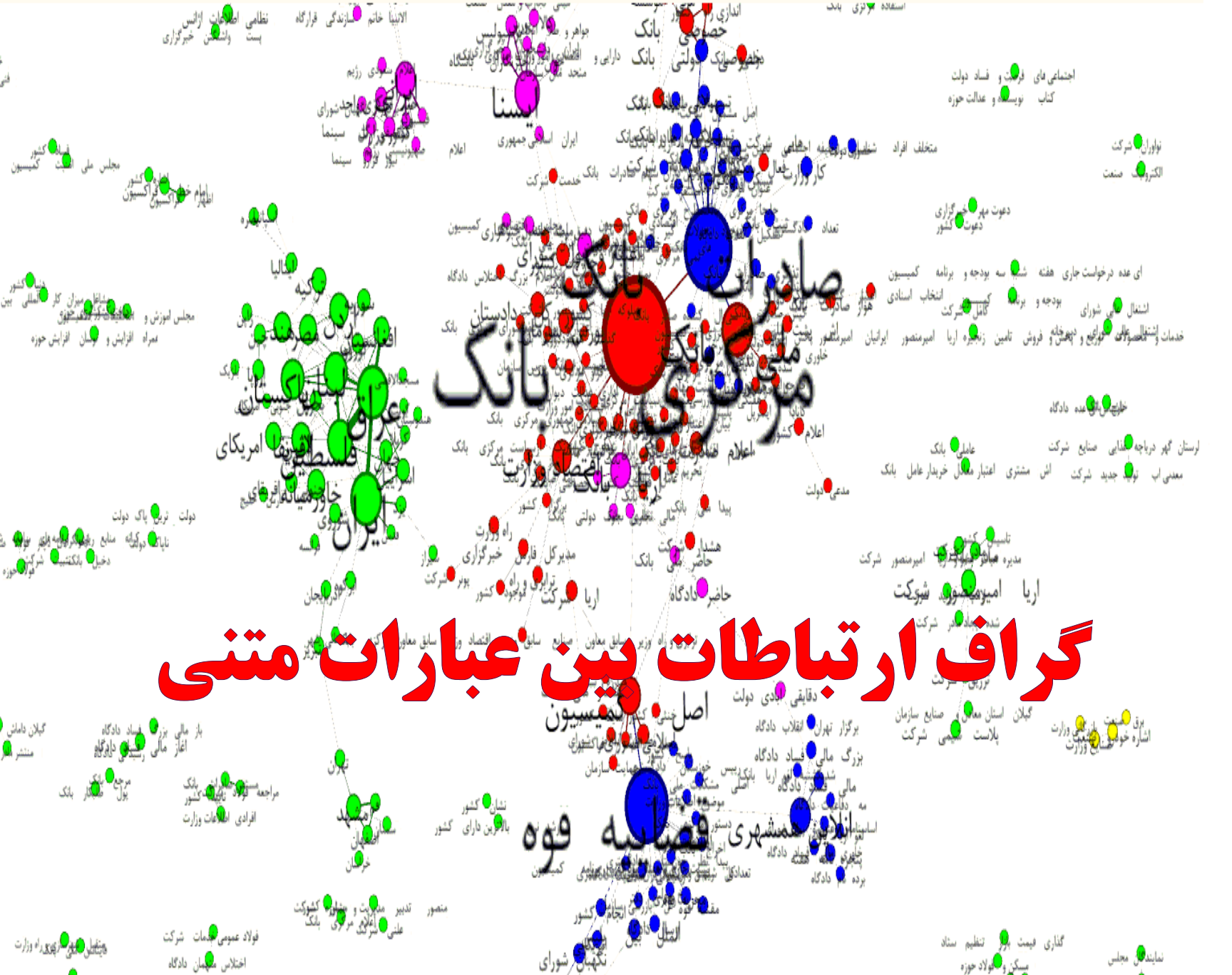


گراف ارتباطات پیامکی (قبل از تحلیل)

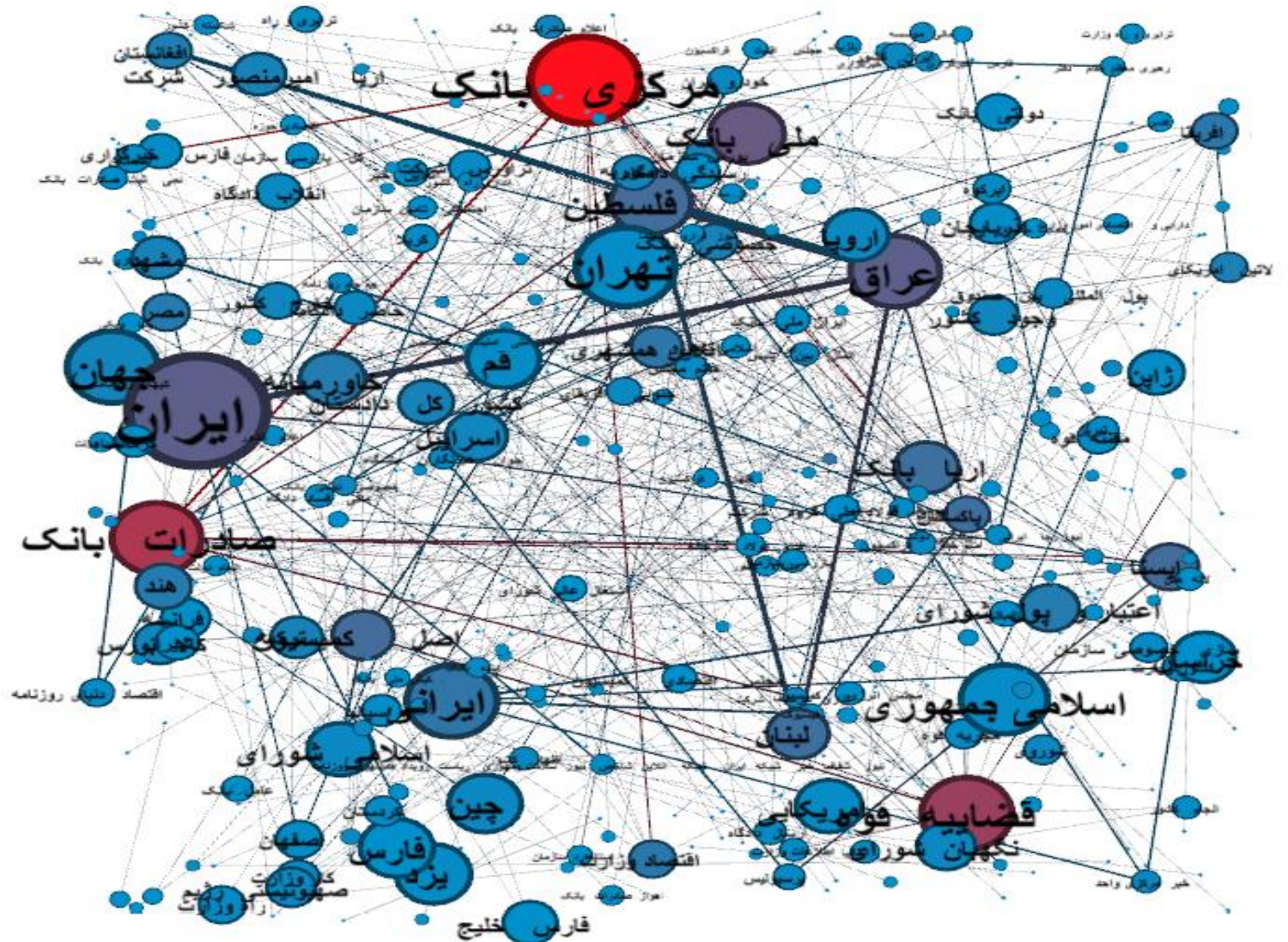


گراف ارتباطات پیامکی

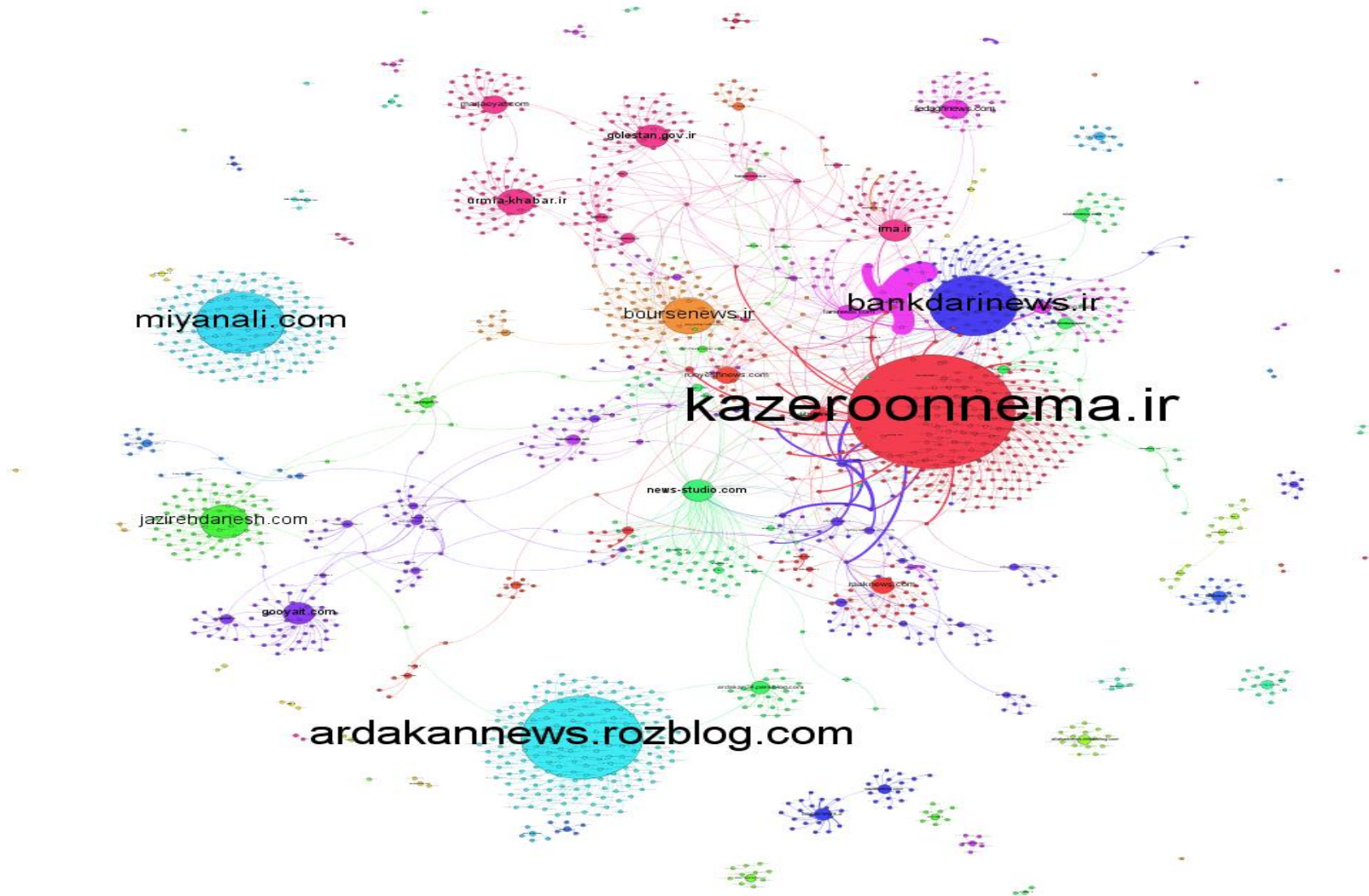




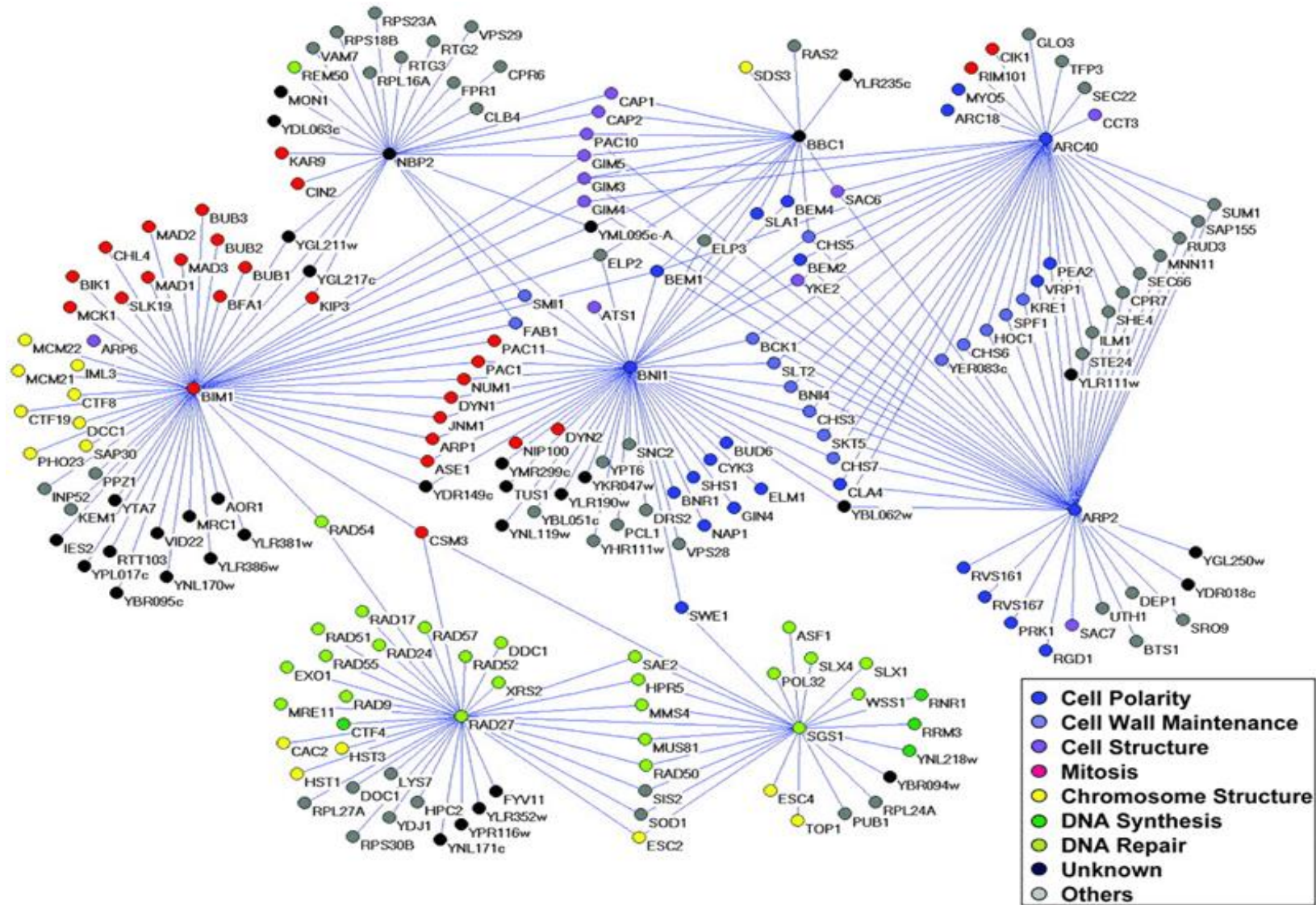
گراف ارتباطات بین موجودیت های اسمی



وب گراف



ارتباطات سلولی و ژنتیکی



ساختار منطقی ذخیره سازی گراف ها

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:
علیرضا حبیبی

حوزه ذخیره و بازیابی گراف

◎ ساختمان داده های منطقی ذخیره سازی گراف

◎ روش ها و تکنیکهای ذخیره سازی گراف

▪ (مثل XML, JSON)

◎ قالب فایل (FILE FORMAT) ذخیره سازی گراف

▪ (مثل GEXF)

◎ پایگاه داده های مبتنی بر گراف

ساختارهای منطقی ذخیره سازی گراف

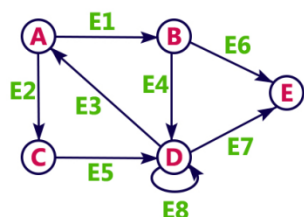
○ ماتریس مجاورتی

○ ماتریس برخورد

○ لیست و آرایه مجاورتی یا وکتور مجاورتی

○ تک جدولی

○ دو جدولی



	E1	E2	E3	E4	E5	E6	E7	E8
A	1	1	-1	0	0	0	0	0
B	-1	0	0	1	0	1	0	0
C	0	-1	0	0	1	0	0	0
D	0	0	1	-1	-1	0	1	1
E	0	0	0	0	0	-1	-1	0

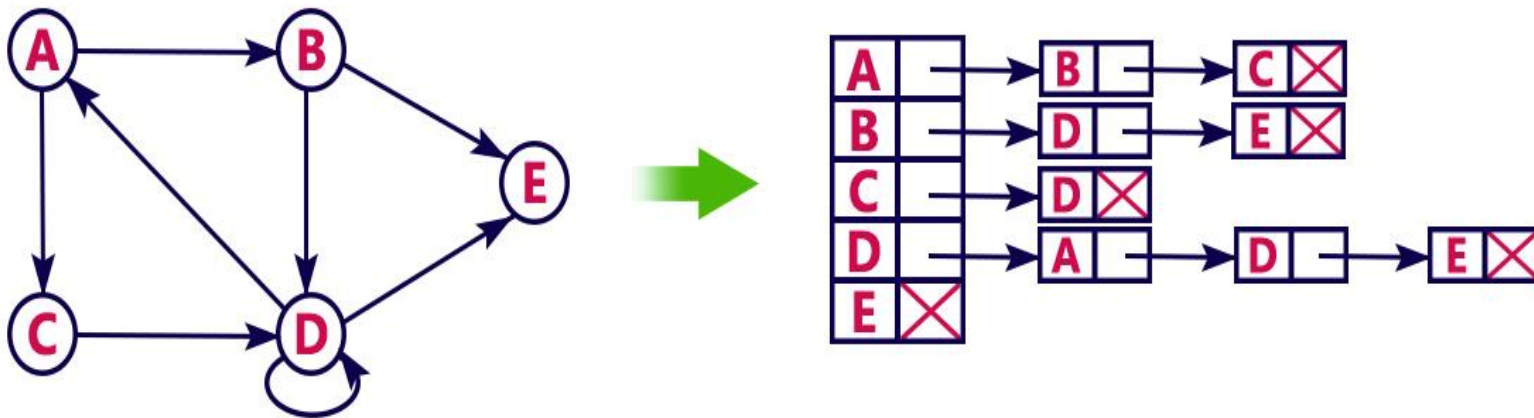
	A	B	C	D	E	F	G	Degree
A	0	1	1	1	1	1	0	5
B	1	0	0	0	0	1	0	2
C	1	0	0	0	0	0	0	1
D	1	0	0	0	0	0	0	1
E	1	0	0	0	0	0	0	1
F	1	1	0	0	0	0	1	3
G	0	0	0	0	0	1	0	1

Adjacency matrices

	A	B	C	D	E	F	G	Out-degree
A	0	1	0	0	0	0	0	1
B	0	0	1	1	0	1	1	4
C	0	0	0	1	0	0	0	1
D	0	0	0	0	1	0	0	1
E	0	0	0	0	0	0	0	0
F	0	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0	0

	A	B	C	D	E	F	G	Degree
A	0	8	12	12	12	16	12	72
B	8	0	0	0	0	4	0	12
C	12	0	0	0	0	0	0	12
D	12	0	0	0	0	0	0	12
E	12	0	0	0	0	0	0	12
F	16	4	0	0	0	0	12	32
G	12	0	0	0	0	12	0	24

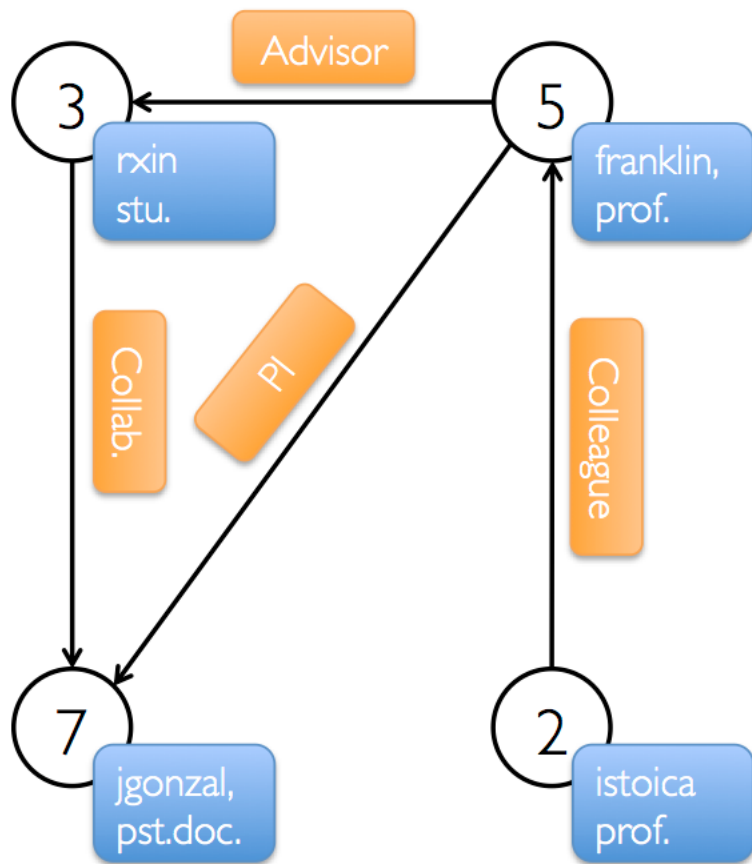
لیست و آرایه مجاورتی یا وکتور مجاورتی



تک جدولی (جدول یال ها)

Source	Target

Property Graph



Vertex Table

Id	Property (V)
3	(rxin, student)
7	(jgonzal, postdoc)
5	(franklin, professor)
2	(istoica, professor)

Edge Table

SrcId	DstId	Property (E)
3	7	Collaborator
5	3	Advisor
2	5	Colleague
5	7	PI

تئوری شبکه های پیچیده پویا

لینک مطلب

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

فهرست مطالب

- تعریف و خصوصیات شبکه های اجتماعی
 - ویژگیهای شبکه های اجتماعی مبتنی بر وب
 - ویژگیهای شبکه های اجتماعی مبتنی بر موبایل
 - انواع شبکه های اجتماعی
 - شبکه های یک نوعی
 - شبکه های دو نوعی
 - شبکه های اجتماعی برخط
 - شبکه های رسانه اجتماعی:
- ارتباطات
 - چگالی
 - مرکزیت
 - کلیک
 - گراف عضویت
 - گراف تصادفی
 - شبکه های پیچیده
 - پدیده دنیای کوچک
 - پل
 - شبکه های مقیاس آزاد
 - جامعه
 - قطر

توپولوژی شبکه

توپولوژی شبکه های اجتماعی نقش مهمی در استخراج برخی اطلاعات از آنها ایفا میکند. توپولوژی شبکه به ساختار و نحوه اتصال نودهای آن اشاره دارد. قسمت اعظم فرایند استخراج توپولوژی شبکه با استفاده از توزیع درجه نودها انجام می شود.

توپولوژی مستقل از مقیاس:

- فاصله متوسط میان نودها بسیار کم است.
- نودهای هاب در این شبکه وجود دارند.
- هاب ها نودهایی هستند که اتصالات زیادی به دیگر اعضا دارند.
- ۸۰ درصد ارتباطات توسط تنها ۲۰ درصد نودها تصاحب می شود.

فراوانی درجات گره ها

DEGREE DISTRIBUTE

- ◉ درجه یک گره تعداد گره هایی است که با آن گره در همسایگی مستقیم قرار دارد.
- ◉ هر چقدر درجه یک گره بیشتر باشد، اهمیت آن گره بیشتر می شود.
- ◉ مثلاً در شبکه وب، اگر درجه ورودی یک دامنه زیاد باشد، یعنی سایت مرجع است، و اگر درجه خروجی آن زیاد باشد، یعنی اینکه سایتی است که از اطلاعات (اخبار) سایت های دیگر استفاده می کند.
- ◉ توزیع درجات، توزیع احتمال این درجات روی کل شبکه است.
- ◉ توزیع درجات شبکه $p(k)$ با استفاده از رابطه $p(k) = \frac{n_k}{n}$ محاسبه می شود.

قطر گراف (DIAMETER)

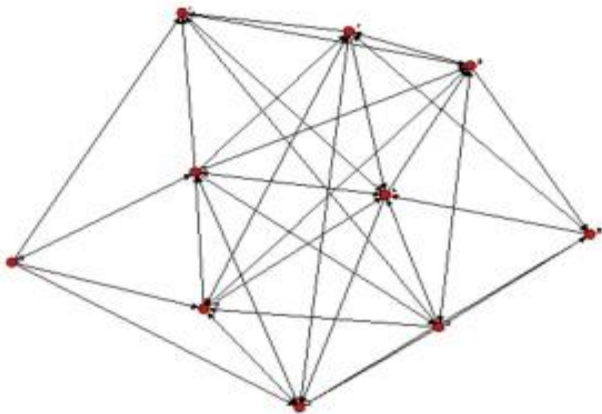
- قطر گراف در واقع بیشترین مقدار از بین کوتاهترین مسیرهای (فاصله) بین گره های گراف است.
- مثلا اگر قطر گراف خبری ۳ بوده است. یعنی هر دو گره حداقل با ۳ گام (یال) به یکدیگر می رسند (در صورتی که بین دو گره مسیری وجود داشته باشد).
- به عبارتی یعنی بزرگ ترین مسیر از بین کوتاه ترین مسیرها بین گره ها
- یعنی فاصله دو گره هرگز از قطر گراف بیشتر نخواهد شد.

Average Degree

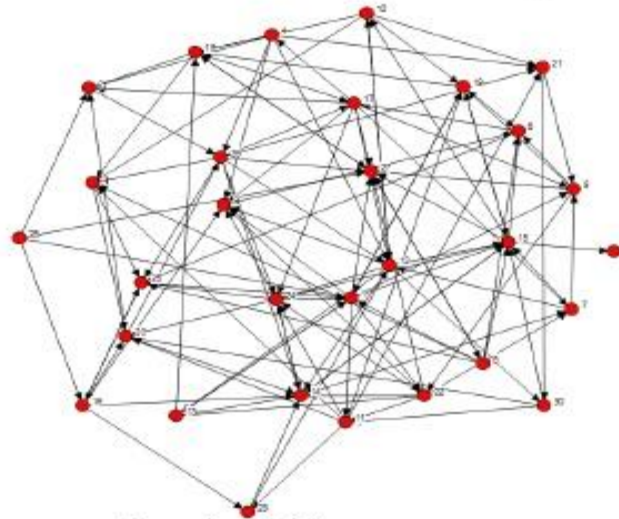
$$\text{Density} = E/N^2$$

$$\text{Directed graph Density} = 2E/N^2$$

- Average number of links per Person



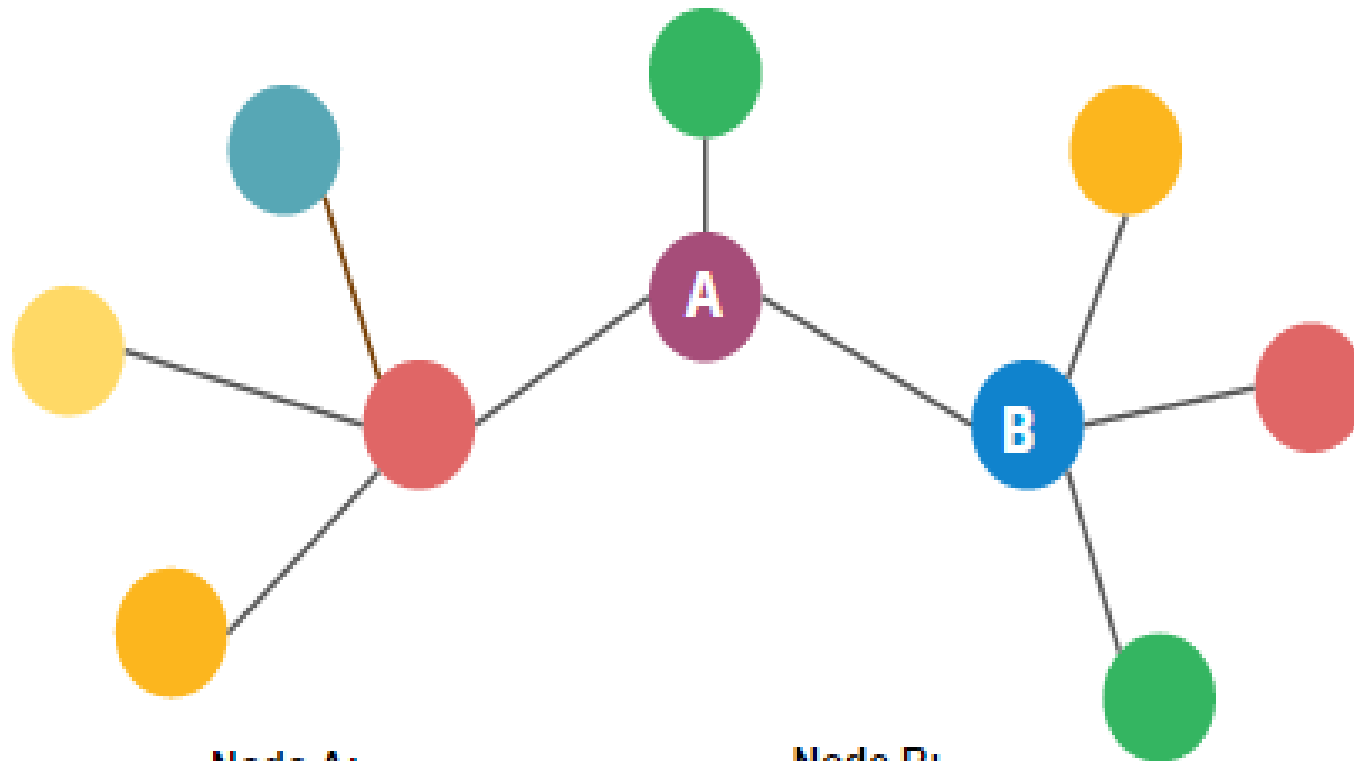
Density: 0,47
Average Degree: 4



Density: 0,14
Average Degree: 4

© Thomas Plotkowiak 2010

مرکزیت بازیگر اصلی در شبکه



Node A:
Eigenvector Centrality: 0.182

Node B:
Eigenvector Centrality: 0.091

انواع مرکزیت

◎ مرکزیت درجه

- در این مرکزیت هر چقدر درجه‌ی یک گره بیشتر باشد، آن گره از اهمیت بیشتری برخوردار خواهد بود.

◎ مرکزیت نزدیکی

- در این معیار مرکزیت، گره‌ای که به سایر گره‌ها نزدیک‌تر باشد از اهمیت بیشتری برخوردار است.

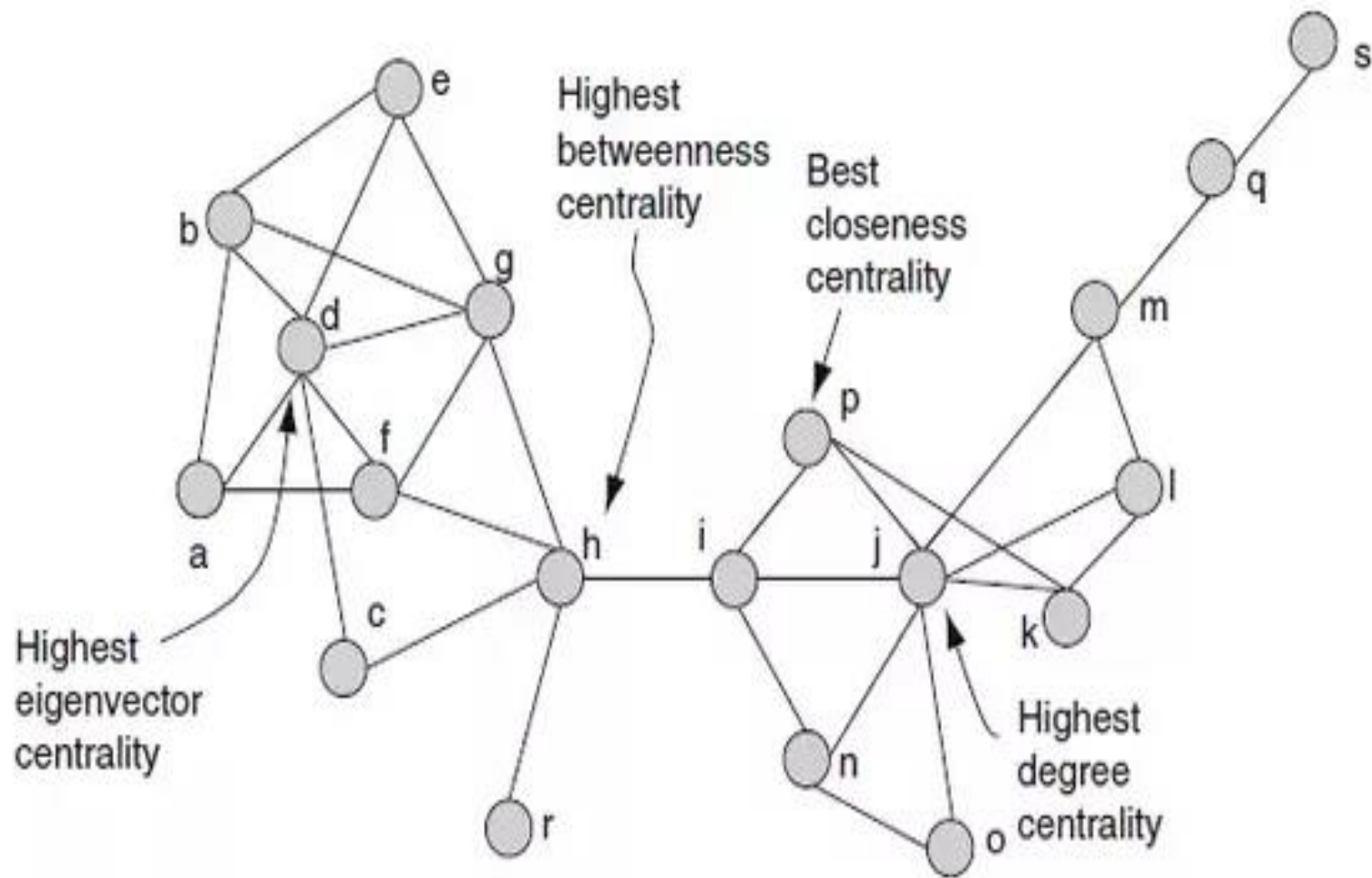
◎ مرکزیت میان‌گذری

- گره‌ای که بر سر راه کوتاه‌ترین مسیر تعداد زیادی از جفت گره‌ها در گراف باشد، با اهمیت‌تر خواهد بود.

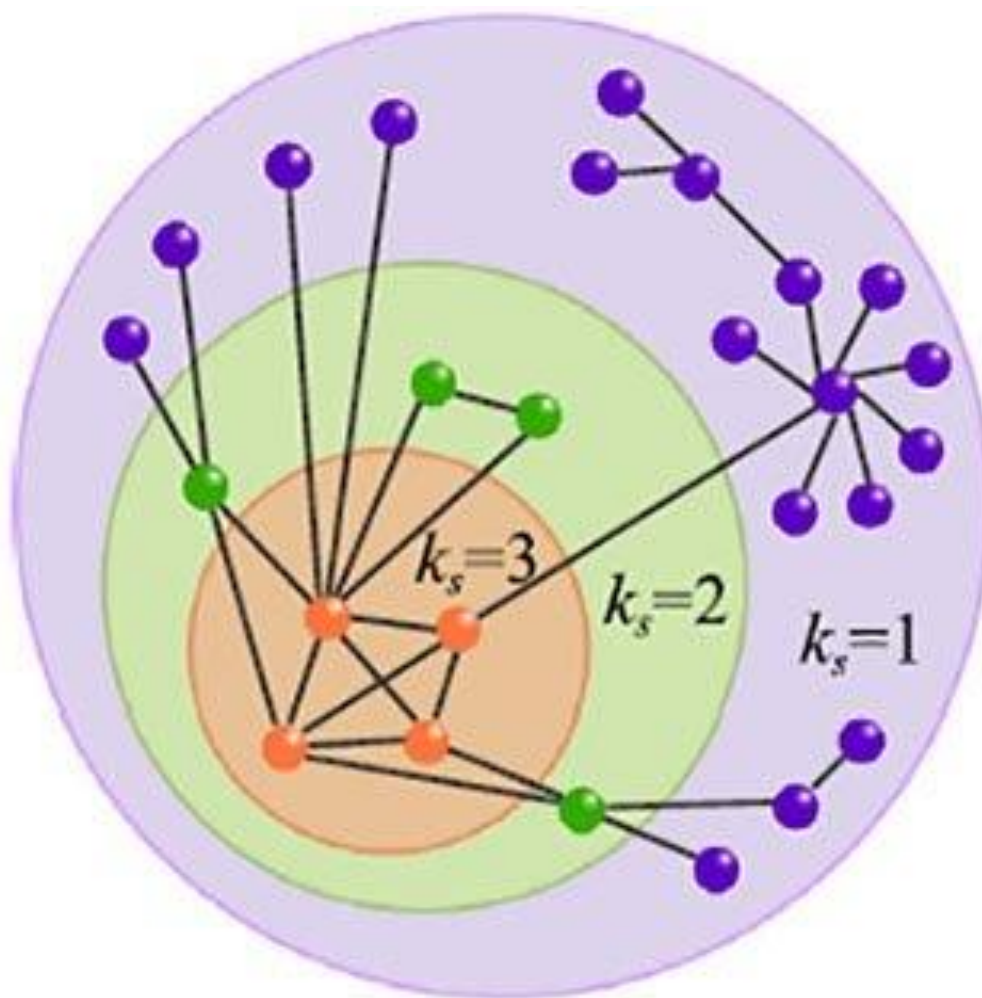
◎ مرکزیت ارزش

- مثل بردار ویژه، پیج رنک گوگل، ضریب پول داران

انواع مرکزیت



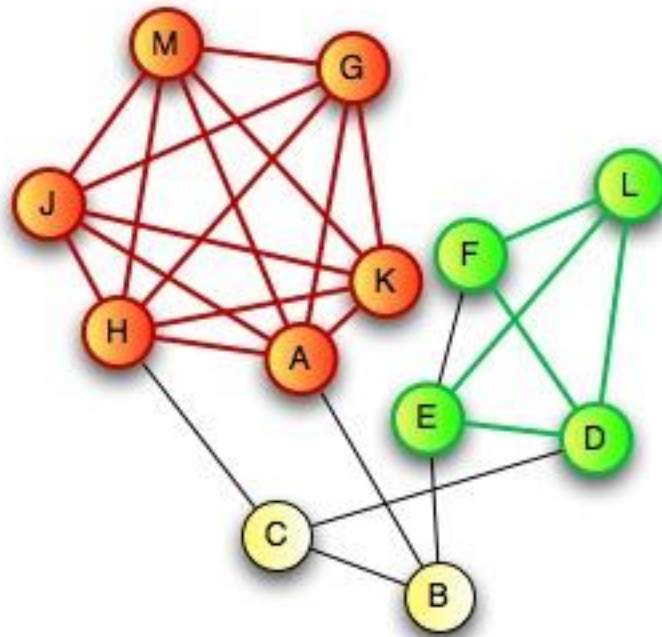
مرکزیت سطح



CLIQUE

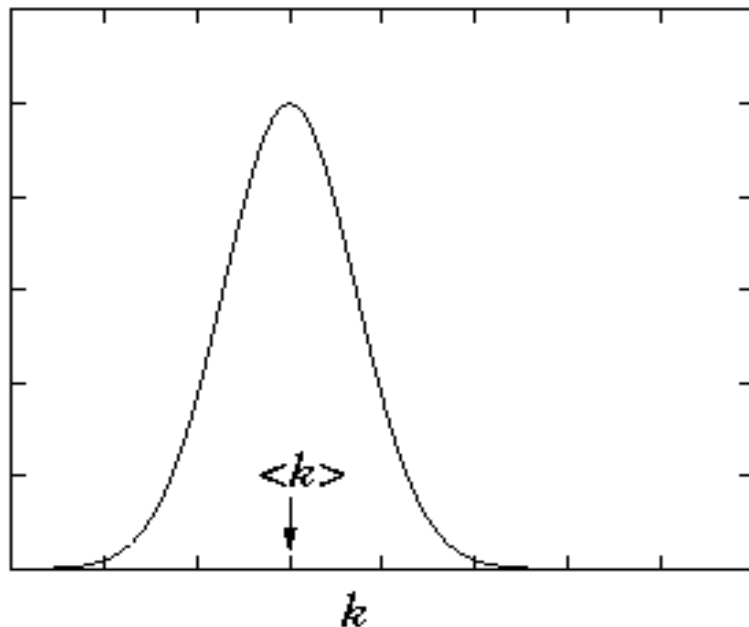
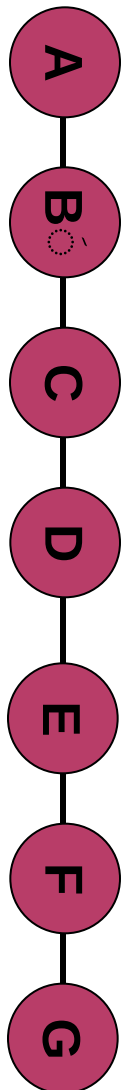
کلیک

⊙ زیر گراف کامل

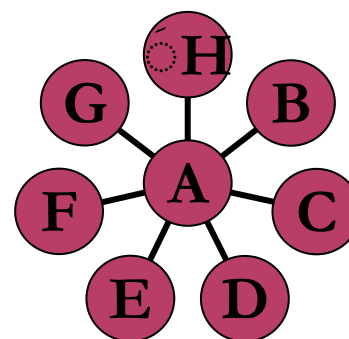
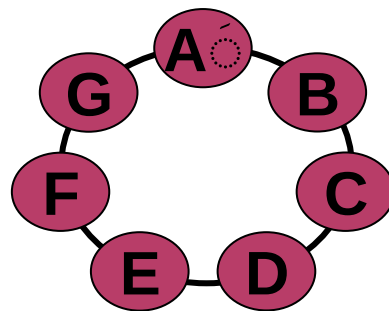


توزیع درجه نودها

فرض اولیه: توزیع پواسون



$$f(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$



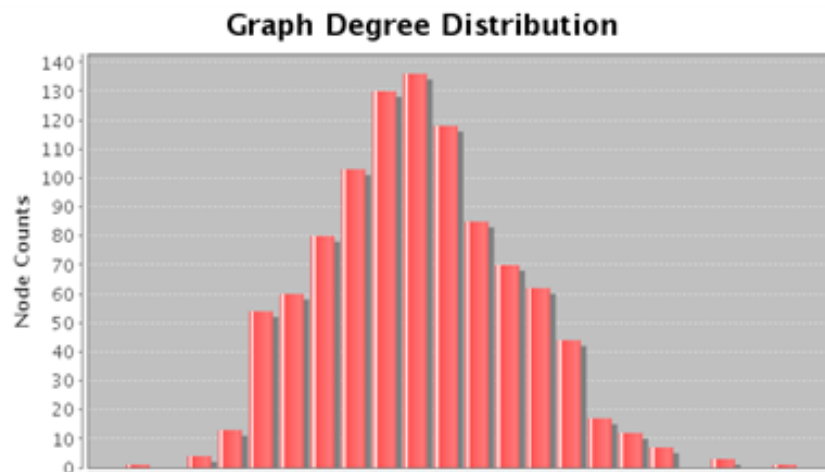
- ◉ دارای قله
- ◉ دارای قرینه
- ◉ توزیع نسبتاً عادلانه لینکها
- ◉ پدیده دنیای کوچک

فراوانی درجات گره ها

DEGREE DISTRIBUTE

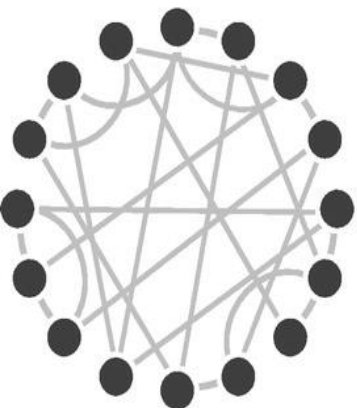
🕒 شبکه های تصادفی

- توزیع درجات شبکه های تصادفی از نوع توزیع پواسون می باشد.
- در این نوع شبکه ها ابتدا تعدادی گره در نظر گرفته می شود و با احتمال p بین هر ترکیب دوتایی از این گره ها یالی افزوده می شود.
- احتمال وجود یال بین هر دو گره از شبکه مقدار یکسانی است.
- شبکه های تصادفی برای مدل سازی شبکه های واقعی استفاده می شوند.

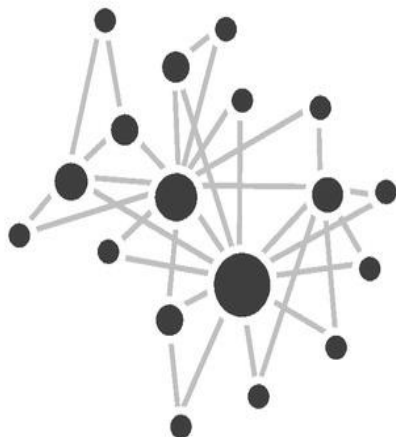


پدیده دنیای کوچک

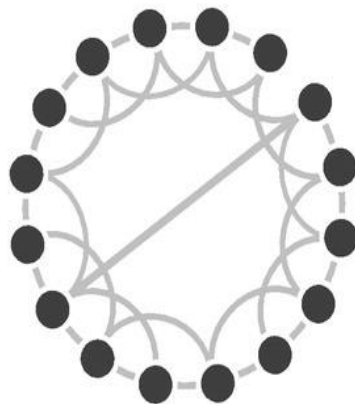
Random graph



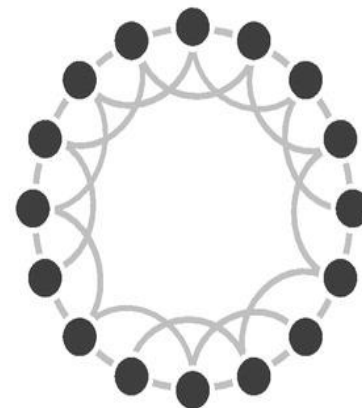
Scale-free network



Small-world network

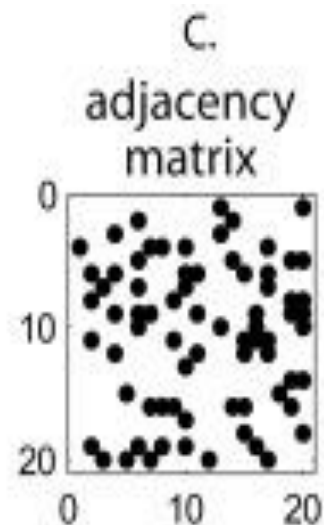
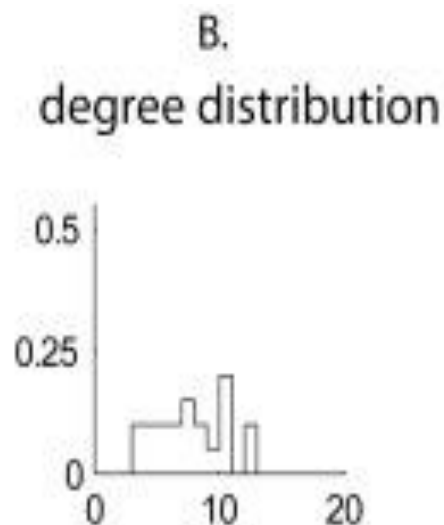


Regular ring lattice



شبکه‌های مقیاس آزاد

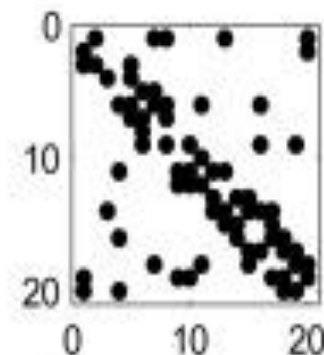
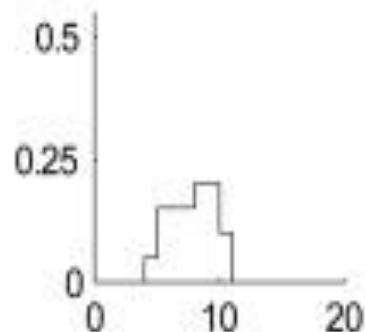
random



D.
Description

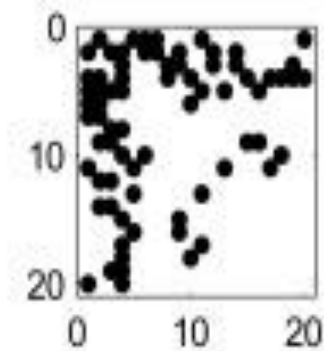
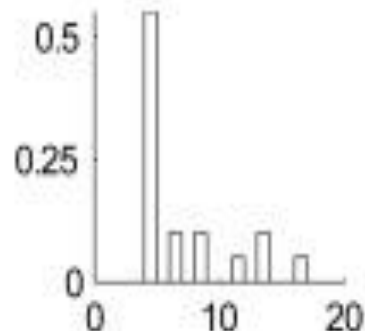
73 connections
among 20 nodes
assigned randomly

small-
world



High local clustering
and short average
path lengths.
'Hub-and-spoke'
architecture.

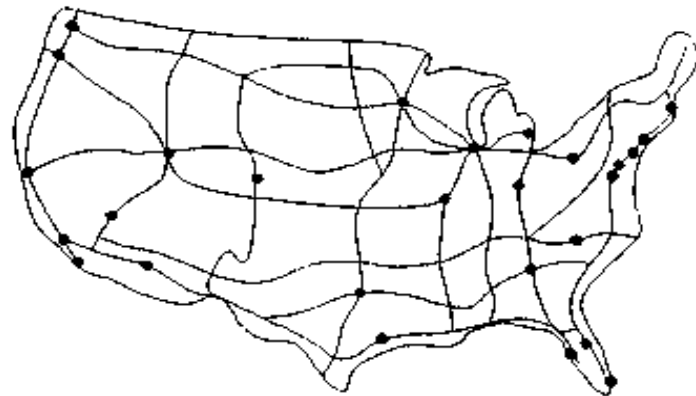
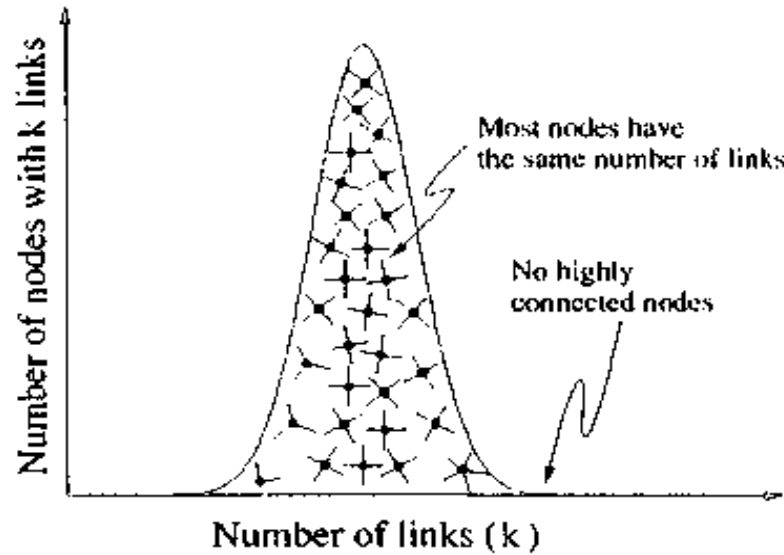
scale-
free



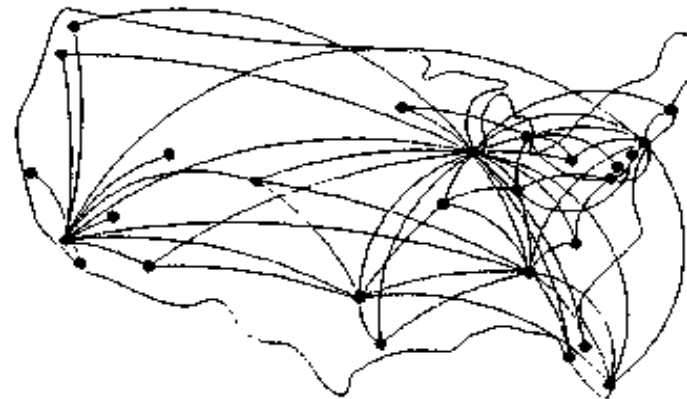
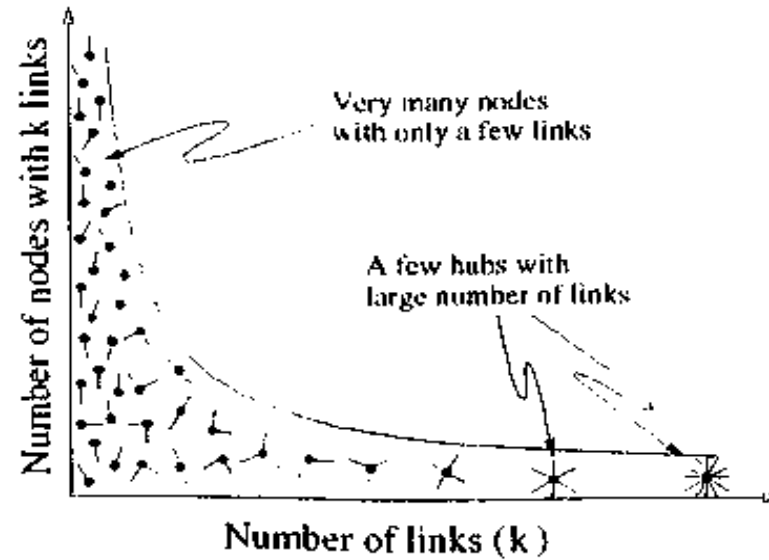
'Hub-and-spoke'
architecture is
maintained at
multiple spatial
scales.

شبکه‌های مقیاس آزاد

Bell Curve



Power Law Distribution



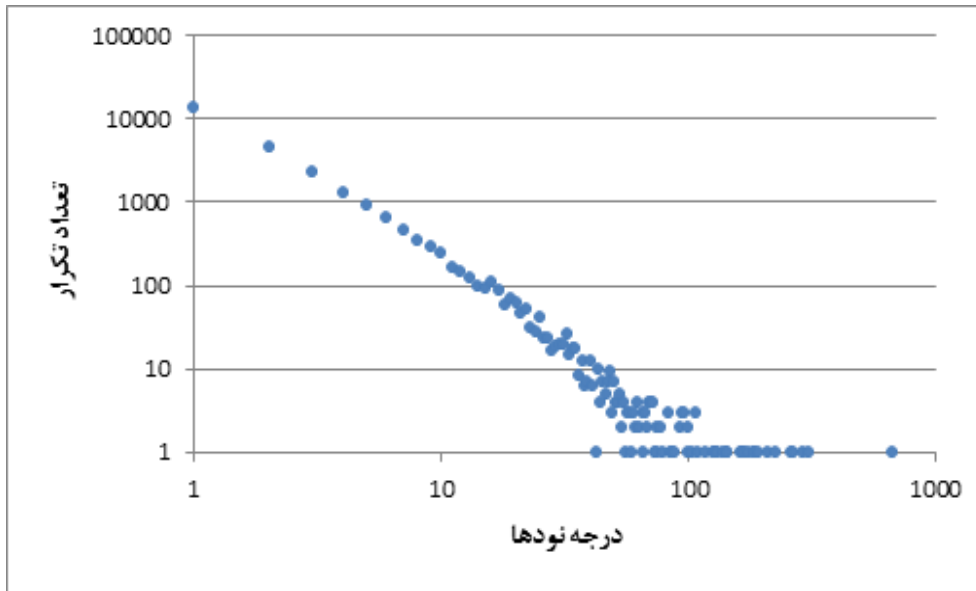
فراوانی درجات گره ها

DEGREE DISTRIBUTE

شبکه واقعی

- توزیع درجات گره ها توزیع توانی (POWRE LAW) می باشد.
- در این نوع شبکه ها تعداد گره هایی با درجات بالا کم و گره هایی با درجات کم زیاد است.
- ساختار این شبکه ها در مقابل خطا بسیار پایدار است. در صورتی که خطایی برای گره های غیر مهم پیش بیاید، خطا به راحتی توسط گره های مهم پوشش داده می شود.
- در صورتیکه برای تعدادی از گره های مهم مشکلی پیش بیاید شبکه کاملاً از هم پاشیده خواهد شد.

- این نوع شبکه ها مستقل از مقیاس نامیده می شوند (خطی).



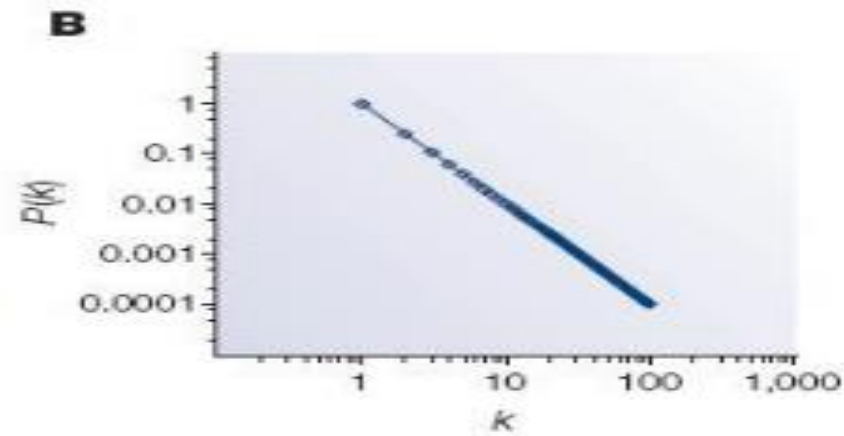
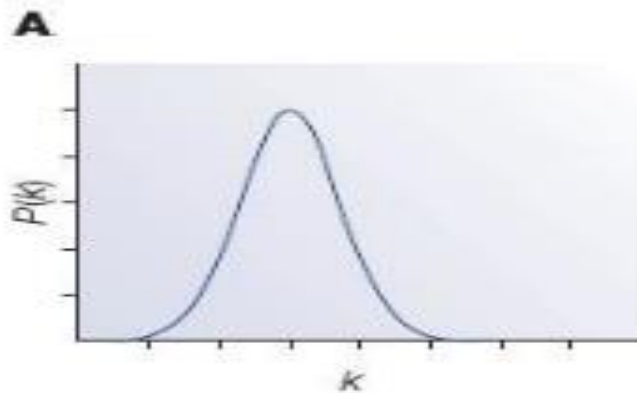
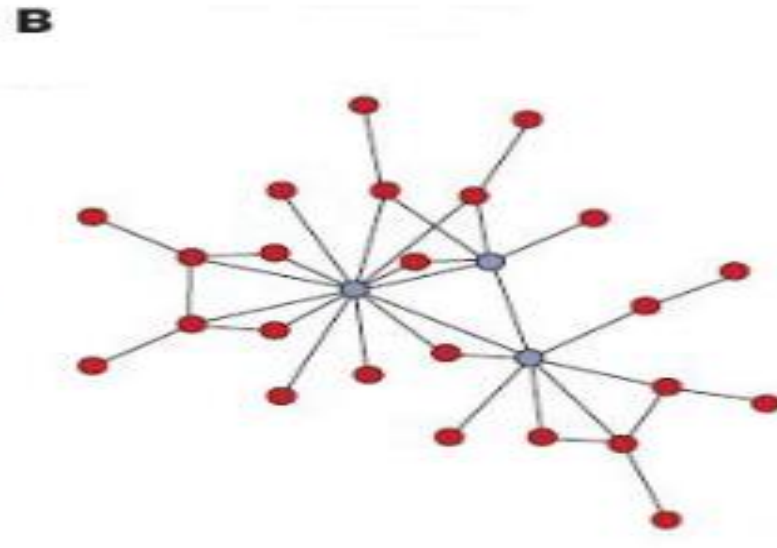
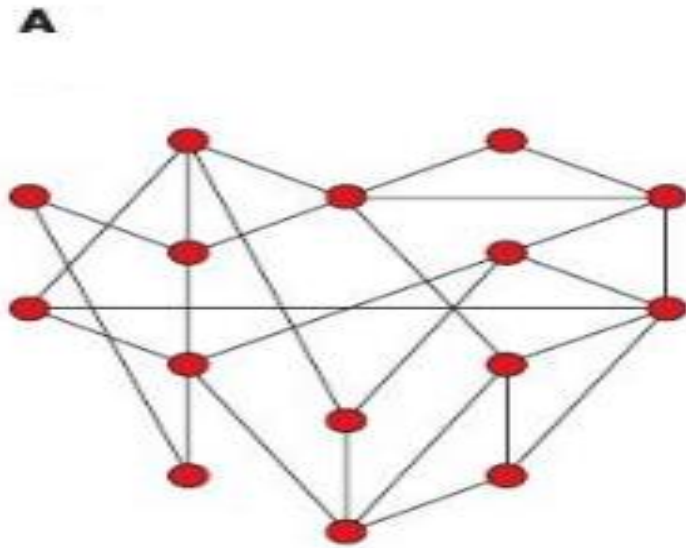
گراف وب

WEB GRAPH

- ◉ گراف وب، گراف جهتداری است که از رابطه بین صفحات (دامنه) بر روی وب بدست می آید.
- ◉ گره ها معرف صفحات (دامنه) هستند.
- ◉ در صورتی که یک لینک بر روی صفحه X وجود داشته باشد که به صفحه Y اشاره می کند. یک یال جهت دار صفحه X را به صفحه Y وصل می کند.
- ◉ توزیع درجات این گراف عموماً مانند توزیع درجات گراف تصادفی Barabási–Albert ، توزیعی توانی است.

گراف وب (WEB GRAPH)

○ A: Random Graph B: Web Graph



قانون PARETO : قانون ۸۰/۲۰

- ۸۰ درصد ثروت دنیا نزد ۲۰ درصد مردم است.
- ۸۰ درصد سود شرکتها ناشی از فعالیت ۲۰ درصد کارمندان است.
- ۸۰ درصد وقت دفاتر خدمات مشتریان صرف ۲۰ درصد از مشتریان می شود.
- ۸۰ درصد تصمیمات در ۲۰ درصد زمان جلسات گرفته می شود.
- ۸۰ درصد پروژه در ۲۰ درصد زمان کل پروژه تمام می شود.
- ۸۰ درصد خلافتها توسط ۲۰ درصد خلافتکاران انجام می شود.

توزیع صحیح: توزیع توانی

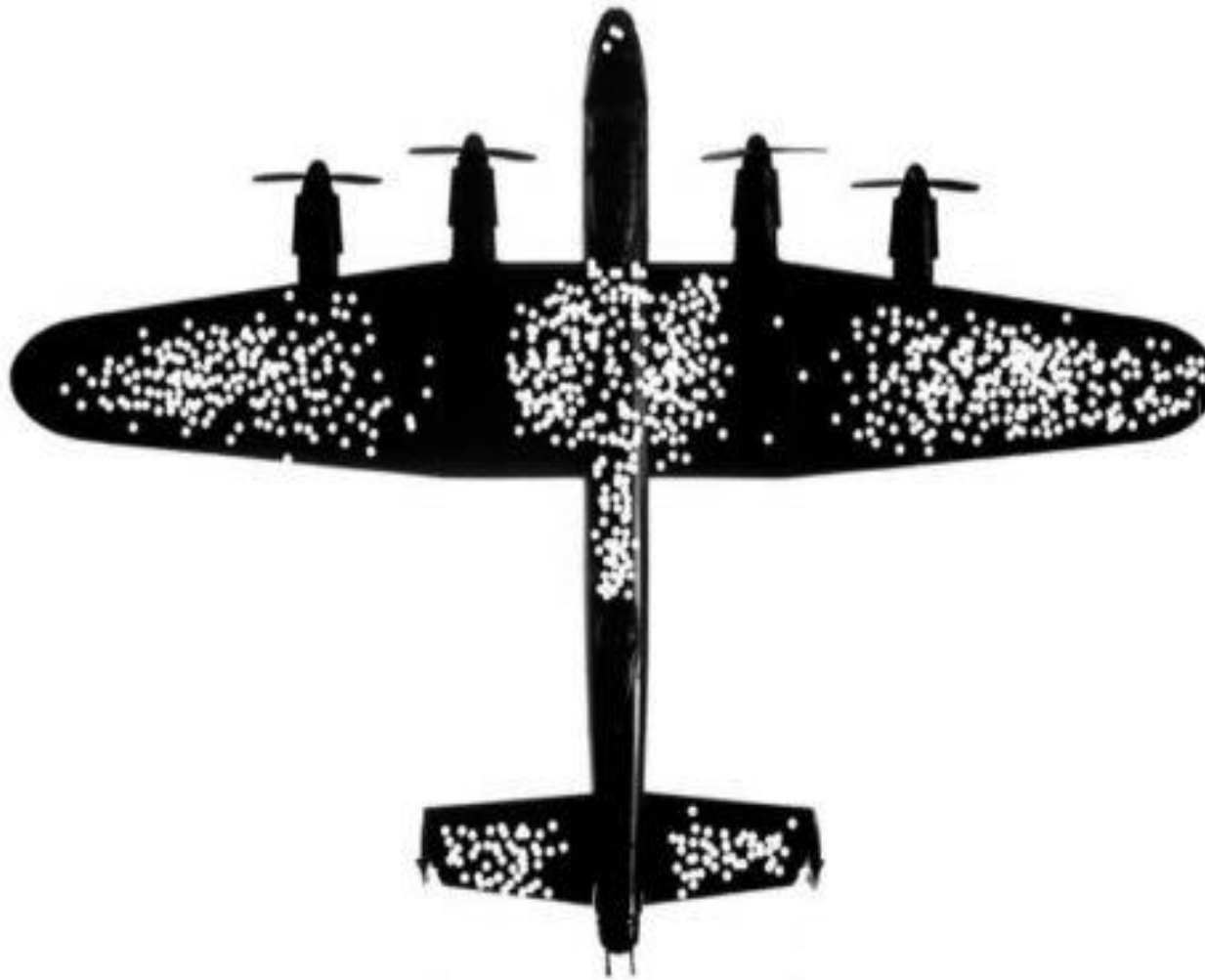


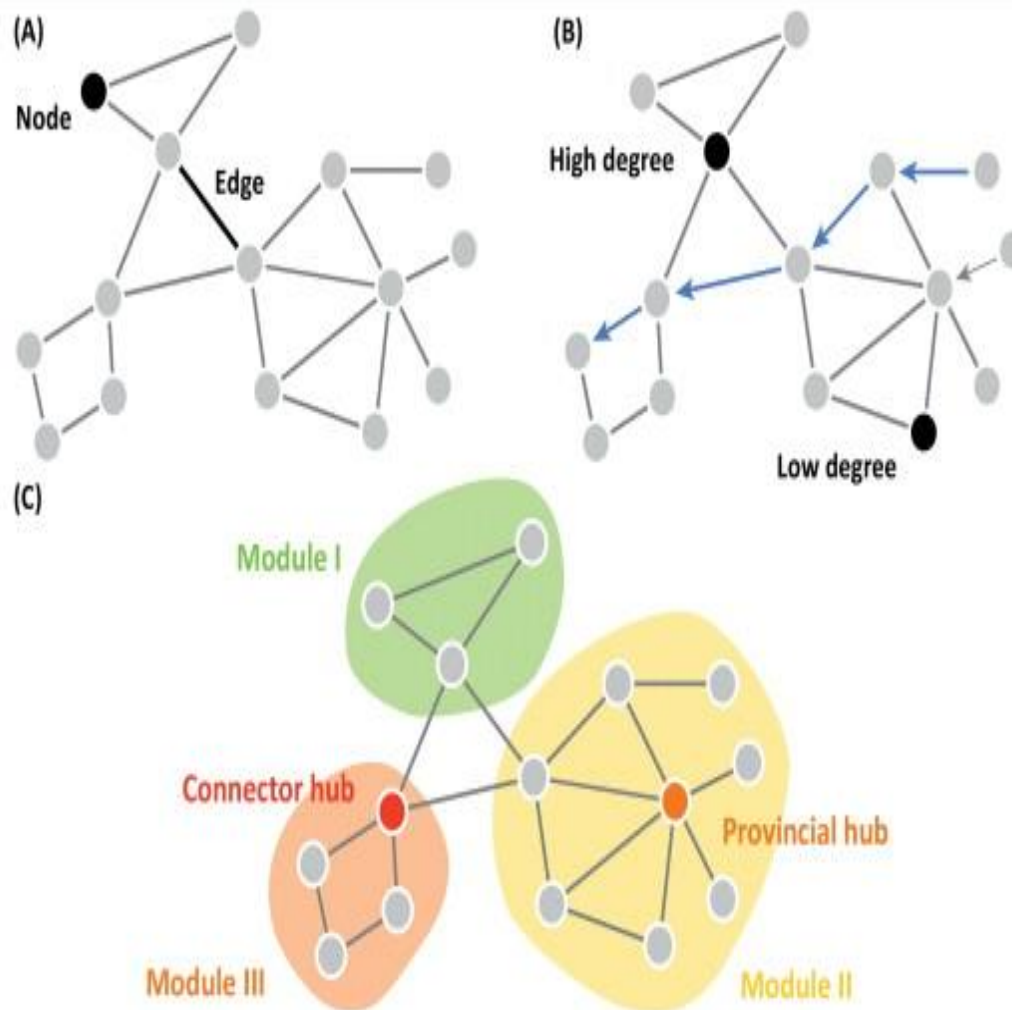
$$p(k) \sim k^{-\gamma}$$

بدون قله
دم بلند
عدم تقارن
مستقل از مقیاس (Scale-free)
ویژگی دنیای کوچک
ویژگی پولدارتر شدن پولدارها

خطای بازماندگان

خطایی در جمع آوری اطلاعات





TRENDS in Cognitive Sciences

قانون PARETO : قانون ۸۰/۲۰

- ① ۸۰ درصد ثروت دنیا نزد ۲۰ درصد مردم است.
- ② ۸۰ درصد سود شرکتها ناشی از فعالیت ۲۰ درصد کارمندان است.
- ③ ۸۰ درصد وقت دفاتر خدمات مشتریان صرف ۲۰ درصد از مشتریان می شود.
- ④ ۸۰ درصد تصمیمات در ۲۰ درصد زمان جلسات گرفته می شود.
- ⑤ ۸۰ درصد پروژه در ۲۰ درصد زمان کل پروژه تمام می شود.
- ⑥ ۸۰ درصد خلافاها توسط ۲۰ درصد خلافاکاران انجام می شود.

◎ حذف تصادفی نودها:

■ آسیب پذیری شبکه‌های مستقل از مقیاس به
خرابی نودها بسیار پایین است!

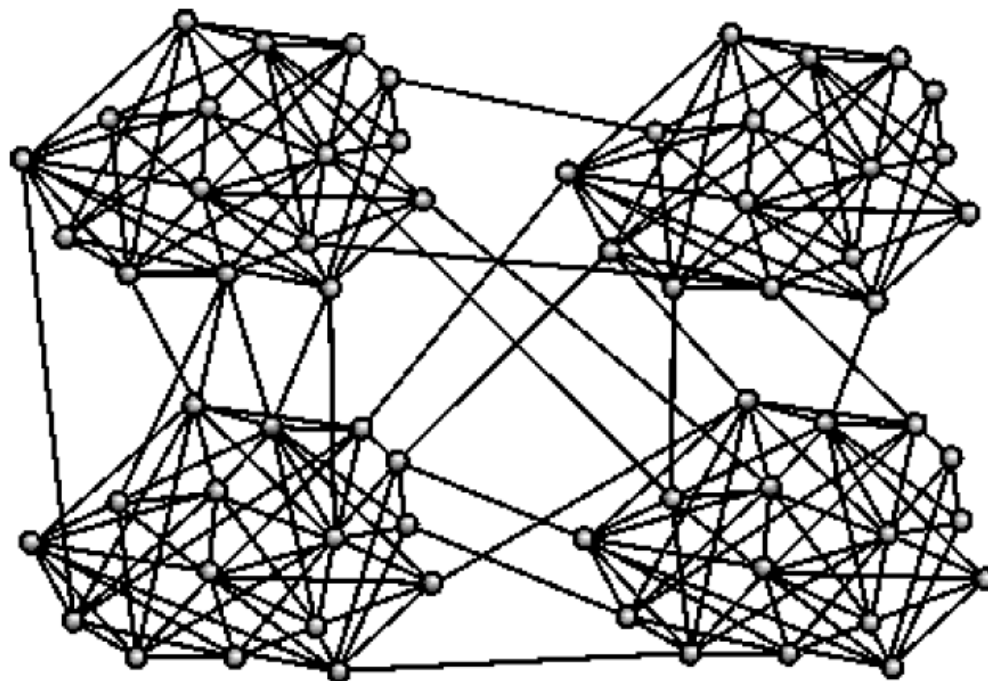
◎ حمله به هاب‌ها:

■ آسیب پذیری شبکه‌های مستقل از مقیاس به
حمله بسیار بالا است!

COMMUNITY DETECTION

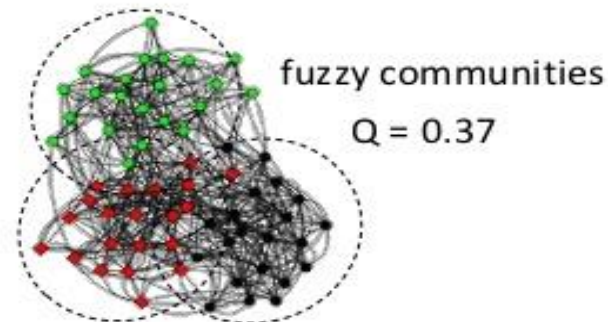
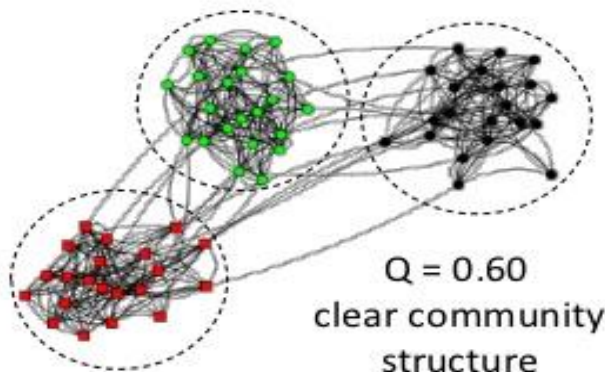
- با روش خوشه بندی
- با روش ماژولاریتی

○ انجمن (Community): گروهی از نودها که دارای ارتباطات درون گروهی نسبتاً زیاد و ارتباطات برون گروهی نسبتاً کم هستند

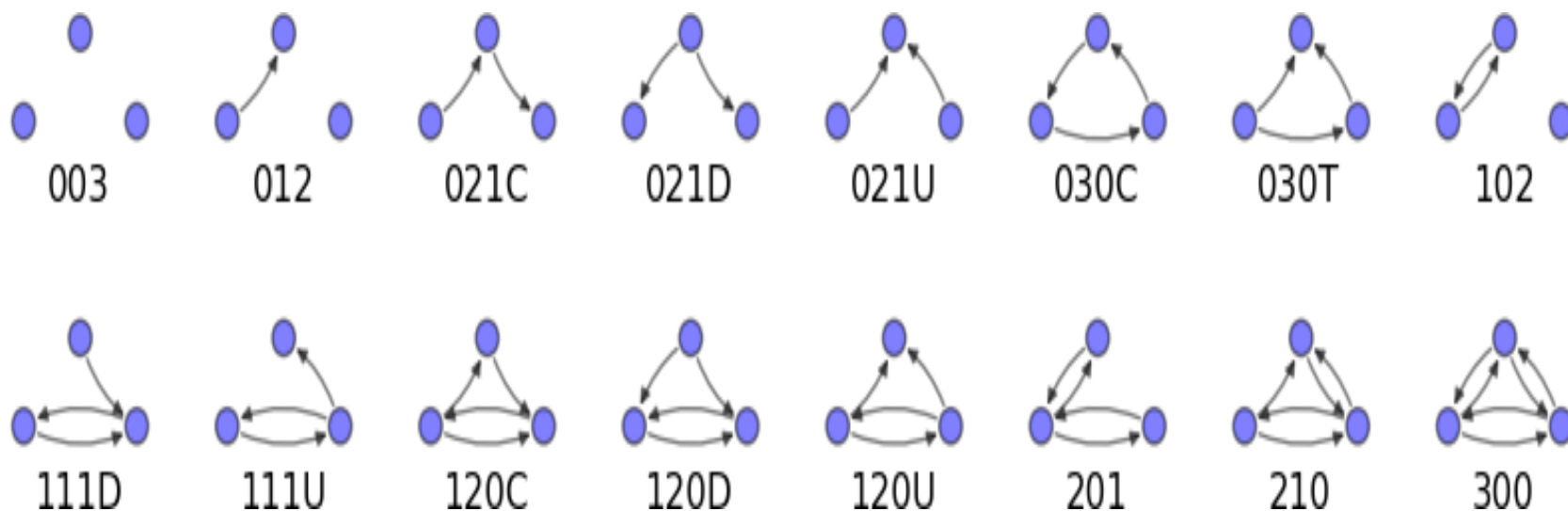


modularity - example

- In a random graph (ER model), we expect that any possible partition would lead to $Q = 0$.
- Typically, in non-random graphs modularity takes values between 0.3 and 0.7.

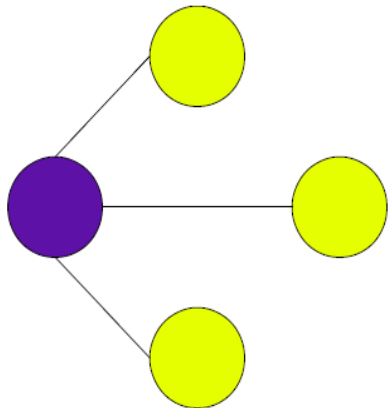


- سال ۱۹۰۸ مهاجرت از زوج های دوتایی به سوی سه تایی ها
- گئورگ زیمل این کار را غنی سازی معنا کرد

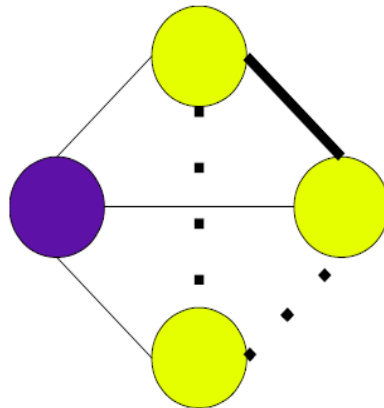


ضرب خوشه بندی

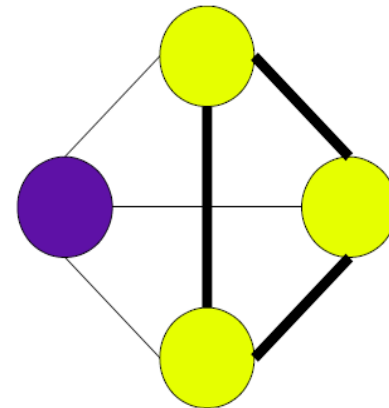
○ احتمال اینکه هر دو نود تصادفی از همسایه ی A با هم دوست باشند. اگر ضرب خوشه بندی زیاد باشد یعنی تراکم بیشتر است.



(a) No pairs formed among neighbors: $C = 0$



(b) One pair formed among neighbors: $C = 1 / 3$



(c) Three pairs formed among neighbors: $C = 3 / 3$

نکات قبل از پردازش گراف

- ◉ جهت دار بودن گراف
- ◉ چگالی گراف
- ◉ توزیع احتمال گراف و تصادفی نبودن
- ◉ بررسی میزان ماژولاریتی ۰.۳ تا ۰.۷
- ◉ عدم خطای بازماندگان در داده ها

GEPHI

نرم افزار تحلیل شبکه های اجتماعی و گراف کاوی

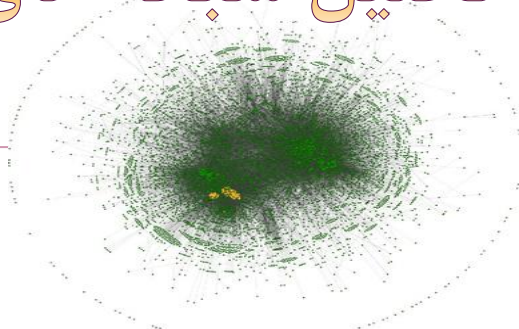
www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

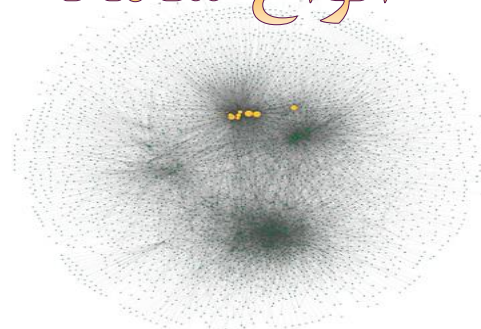
مدرس:
علیرضا حبیبی

انواع سامانه های تحلیل شبکه های اجتماعی

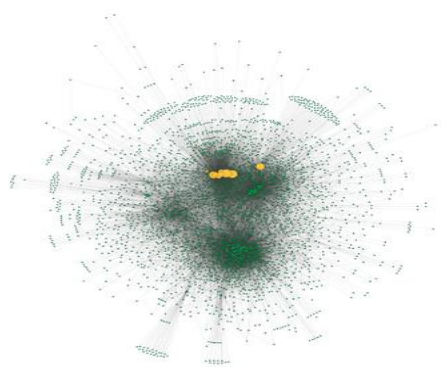
- Nood tools
- Cytoscape
- Gephi



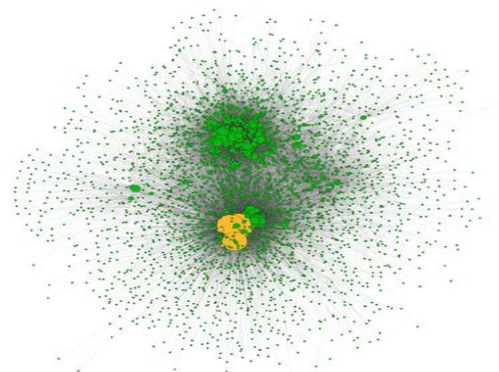
Gephi – Fruchterman Reingold
large network



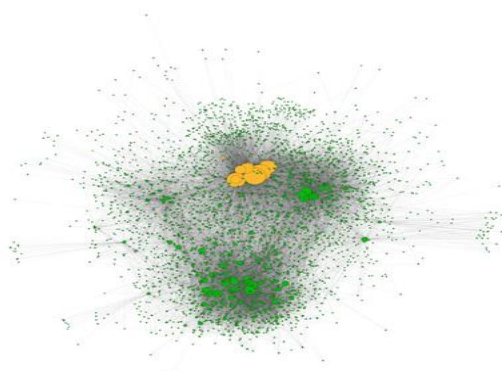
Gephi – Fruchterman Reingold
1st neighbor



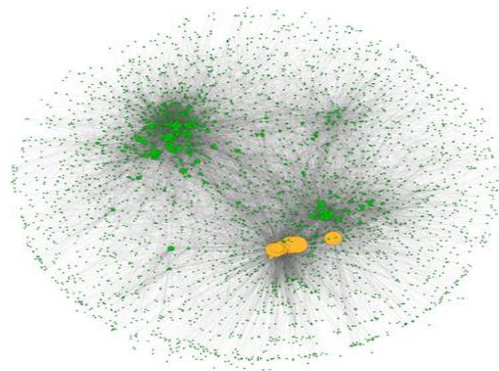
Gephi – Yifan Hu - 1st neighbor



Cytoscape – Force - 1st neighbor



Cytoscape – Organic - 1st neighbor



Cytoscape – Spring - 1st neighbor

ویژگی های گفی

- گفی با استفاده از الگوریتم های بصری سازی در زمان واقعی (Realtime) گره ها در فضای دو یا سه بعدی مستقر میکند.
- با MOVE کردن گره ها بر هم تاثیر متقابل می گذارد.
- گفی از ویژگی گره ها برای تغییر رنگ و اندازه گره ها استفاده می کنند و گروه ها و اشخاص را پیدا می کنند.
- الگوریتم های مختلفی برای تحلیل و نمایش شبکه بر روی آن پیاده سازی شده
- تعداد زیادی برنامه نویس بر روی آن در حال کار هستند.
- این برنامه امکان اضافه شدن قابلیت های را به صورت پلاگین (افزونه) دارد
- قابلیت های نمایش بصری آن به نسبت از بقیه نرم افزارهای موجود بهتر است
- قادر به تحلیل و نمایش شبکه های ایستا و پویا است.
- ابزارهای آماری این نرم افزار کمک شایانی به استخراج خواص آماری مختلف گراف ها می کند.

- ◉ به طور خلاصه گفی یک ابزار بصری سازی برای سیستم عامل های
Windows Mac OS X, Linux
- ◉ به زبان برنامه نویسی جاوا 1.8 و توسعه شده با فریم ورک
Netbeans
- ◉ دارای استاندارد متن باز 3 GNU GPL + CDDL است.
- ◉ معماری گفی مبتنی بر پلاگین است. و کاربران می توانند
کارکردهای مطلوب خود را در به صورت افزونه به آن اضافه کنند.
- ◉ گراف آزمایشی

■ <http://gephi.org/datasets/LesMiserables.gexf>

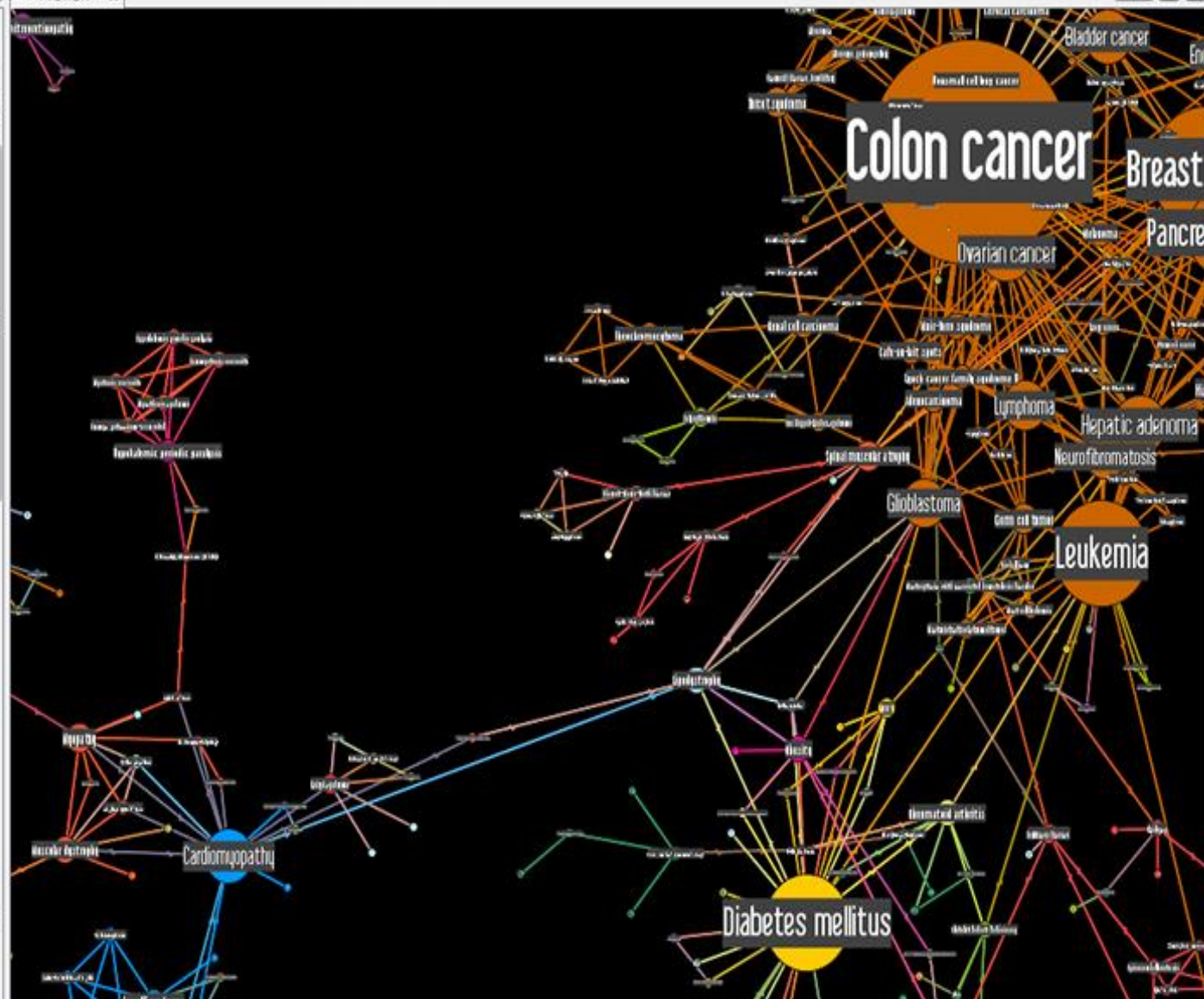
- Show ☒
- Border width 1.0
- Color original
- Border color custom [0,0,0]
- Show labels ☒
- Label font Libel Serif 24 Plain
- Shorten labels ☐
- Shorten limit 10
- Label color custom [255,255,255]
- Label border ☒
- Label Border Color custom [66,66,66]
- Edge
- Show ☒
- Self-Loop
- Show ☒
- Thickness 1.0
- Color custom [0,0,0]
- Undirected
- Curved ☐
- Thickness 1.0
- Color mixed
- Show labels ☐
- Shorten labels ☐
- Shorten limit 10
- Font SansSerif 10 Plain
- Label color parent
- Directed
- Curved ☒
- Thickness 1.0
- Color source

Preview ratio: 100 %



Refresh

Export: SVG



Background Reset zoom



Presets

Default

Shorten labels ☐

Shorten limit 10

Font SansSerif 10 Plain

Label color parent

Directed

Curved ☒

Thickness 0.01

Color source

Labels ☐

Shorten labels ☐

Shorten limit 10

Font SansSerif 10 Plain

Label color parent

Mini-Labels ☐

Mini-Label radius 15.0

Shorten Mini-Labels ☐

Mini-Label limit 10

Mini-Label font SansSerif 8 Plain

Mini-Label color parent

Arrows ☒

Arrow added radius 65.0

Arrow size 20.0

Arrow color parent

Mutual

Curved ☒

Thickness 0.5

Color mixed

Labels ☐

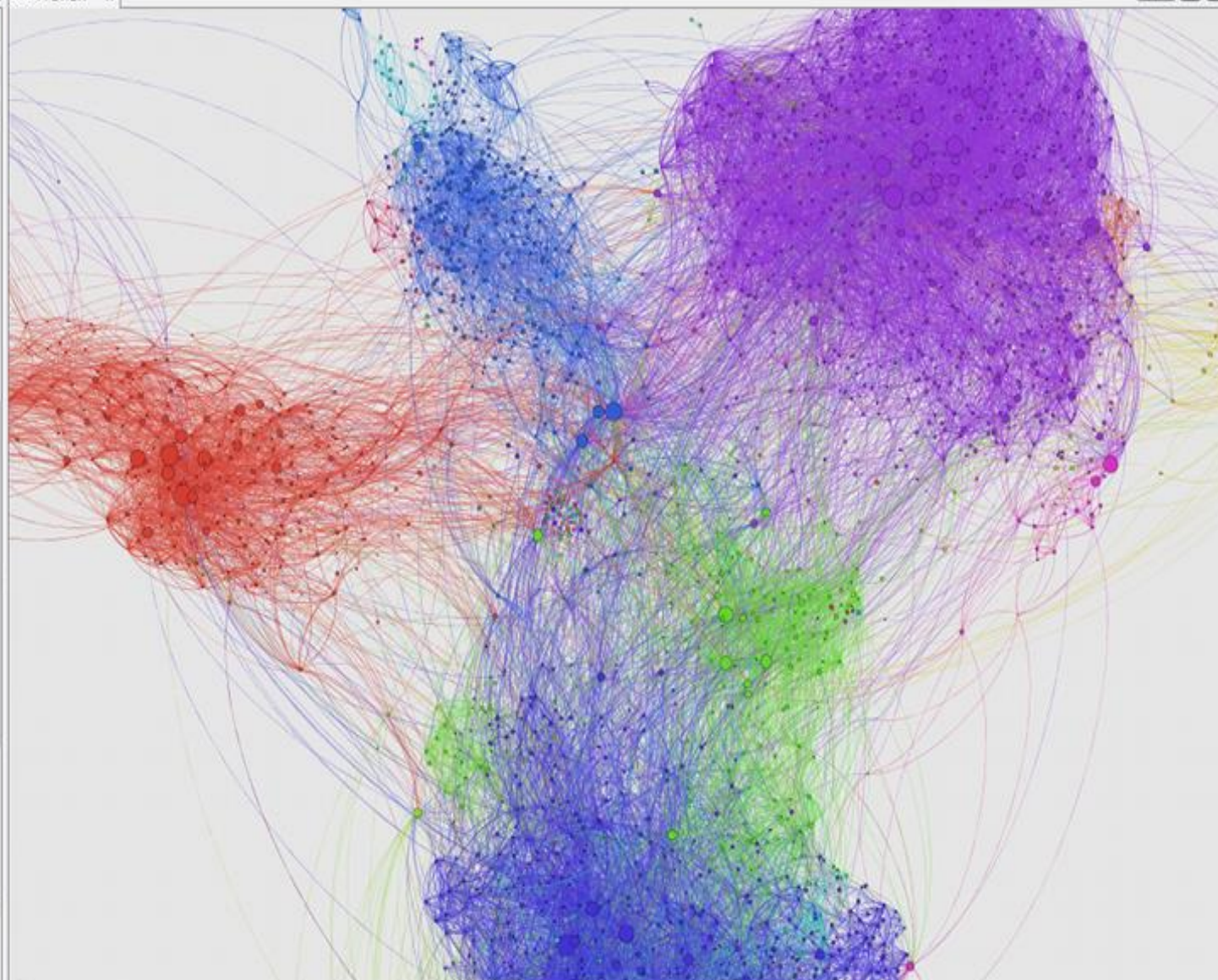
Shorten labels ☐

Shorten limit 10

Preview ratio: 100 %



Export: SVG

☐ Background Reset zoom

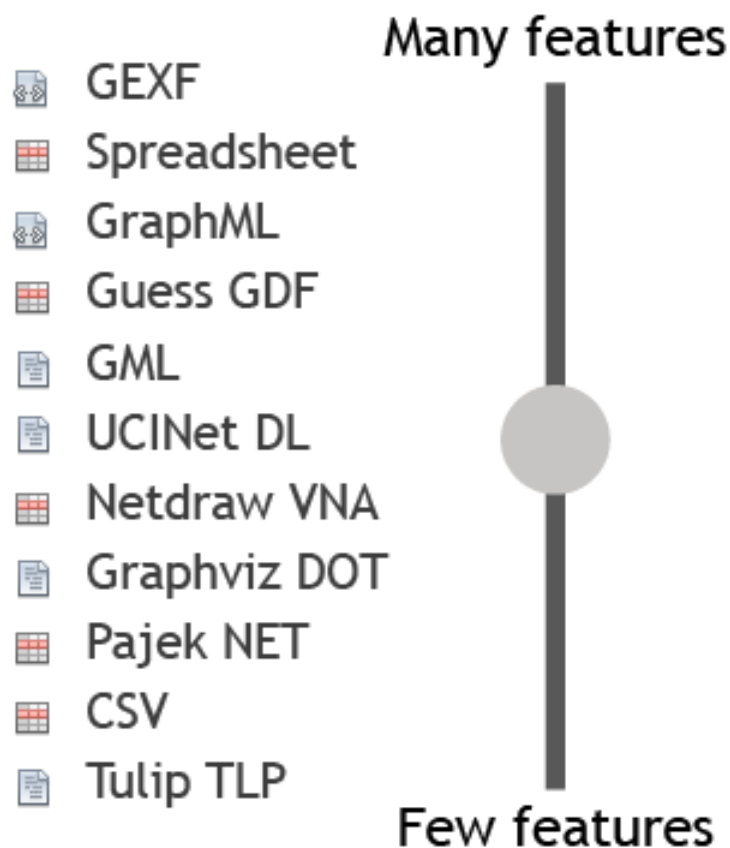
فرمت فایل های ورودی برای گفی

- ابزار گفی عمده ی فرمت های معمول و معروف گراف را پشتیبانی میکند. در زیر برخی از فرمت های قابل پشتیبانی توسط گفی آورده شده است.
- البته فرمت فایل خاص این نرم افزار پسوند `gephi` دارد.

- CSV
- Pajek NET
- Guess GDF
- GEXF
- GraphML
- Graphviz DOT
- UCInet DL
- NetdrawVNA
- Tulip TLP
- Excel Spreadsheetater
- ...

	Edge List/Matrix Structure	XML Structure	Edge Weight	Attributes	Visualization Attributes	Attribute Default Value	Hierarchical Graphs	Dynamics
CSV								
DL Ucinet								
DOT Graphviz								
GDF								
GEXF								
GML								
GraphML								
NET Pajek								
TLP Tulip								
VNA Netdraw								
Spreadsheet*								

Do you need...



File Type

-  XML
-  Tabular
-  Text

Gephi requirements

Gephi is a cross-platform application, as it is developed in Java. It was successfully tested on many different architecture, OS and graphical configuration. Gephi requires Java version 7 and later, get it [here](#).

Minimum hardware requirements

500 MHz CPU + 128 MB RAM + OpenGL 1.2

There is no absolute minimum. This is a practical minimum. Performance is a function of graph size.

Network size (nodes + edges)	~Memory suggested
~1000	128mo
~10,000	512mo
~100,000	2go
~1M	>8go

Note that Gephi has been successfully tested on Netbooks, running on Intel® Atom.

Graphics

Gephi uses an OpenGL 3D engine to speed up graph visualization. However a compatible graphic card is required. If your graphic card is older than 8 years, or if your laptop doesn't have a dedicated graphic card, you may have to upgrade your hardware to run Gephi.

- [Detect and fix graphical problems](#)

If you're running on Linux, get latest vendor graphical drivers to ensure OpenGL hardware acceleration.

Java

The current stable version of Gephi will only run with Java 7 or 8. On Mac OS X, Java is bundled with the application so it doesn't have to be installed separately. On Windows and Linux, the system must be equipped with Java.

CSV Format

Gephi supports CSV files that simply represents relationships. The CSV format can be obtained from any row data, databases or Excel export. Each line must contains at least two elements, separated by a separator (**coma, semicolon, pipe or whitespace**). Values can be encapsulated by single or double quotes. By default graphs imported from CSV are directed graphs, but the user can select undirected in the import report dialog.

Edge List

The CSV example below represents a graph with two edges: "a" -> "b" and "b" -> "c".

```
a;b
b;c
```

Adjacency List

All edges can be written as node pairs. It's also possible to write all node's connection on the same line. The example below represents a graph with 3 edges: "a" -> "b", "b" -> "c" and "b" -> "d"

```
a;b
b;c;d
```

Mixed

The following example shows various cases that CSV supports as well. Self-loops and mutual edges are supported. It's also possible to repeat an edge, "D" -> "E" is repeated twice in this example. **As a consequence the edge weight is incremented.** "D" -> "E" has a edge weight at two, whereas default value is one.

```
A,B
B,A
C,C
D,E
A,D
D,B,E
E,G,A,B
```

Matrix

The sample below shows a graph with 5 nodes. An edge is created when the the cell is '1'.

```
;A;B;C;D;E
A;0;1;0;1;0
B;1;0;0;0;0
C;0;0;1;0;0
D;0;1;0;1;0
E;0;0;0;0;0
```

Edge weight

Simply replace '1' values by the edge weight value, formatted as a 'double'.

NODE CSV

- Id,Label,timeset,modularity_class
 - 9321,__ali__,"<[2008.0, 2012.0]>",82
 - 9322,_amir_,"<[2007.0, 2012.0]>",81
 - 9323,_mosafer_,"<[2010.0, 2012.0]>",35
 - 9324,_sd_,"<[2007.0, 2012.0]>",81
 - 9325,-senjed-,"<[2007.0, 2012.0]>",81
 - 9326,00iran,"<[2009.0, 2012.0]>",42
 - 9327,0soghrat0,"<[2007.0, 2012.0]>",81

Edge CSV

Source,Target,Type,Id,Label,timeset,Weight

9323,10126,Directed,20237,, "<[2010.0, 2011.0]; [2011.0, 2012.0]>",1

9324,9365,Directed,20238,, "<[2007.0, 2008.0]; [2008.0, 2009.0]; [2009.0, 2010.0]; [2010.0, 2011.0]; [2011.0, 2012.0]>",1

9324,9731,Directed,20239,, "<[2007.0, 2008.0]; [2008.0, 2009.0]; [2009.0, 2010.0]; [2010.0, 2011.0]; [2011.0, 2012.0]>",1

9324,9770,Directed,20240,, "<[2007.0, 2008.0]; [2008.0, 2009.0]; [2009.0, 2010.0]; [2010.0, 2011.0]; [2011.0, 2012.0]>",1

9324,10803,Directed,20241,, "<[2007.0, 2008.0]; [2008.0, 2009.0]; [2009.0, 2010.0]; [2010.0, 2011.0]; [2011.0, 2012.0]>",1

فرمت فایل های خروجی برای گفی

◉ نرم افزار گفی برخی از فرمت های معمول گراف را به عنوان خروجی تولید کند. در زیر برخی از فرمت های قابل پشتیبانی توسط گفی آورده شده است.

■ CSV

■ PAJEK NET

■ GUESS GDF

■ GEXF(Graph Exchange XML Format)

■ GRAPHML

■ EXCEL SPREADSHEET

■ SVG

■ PDF

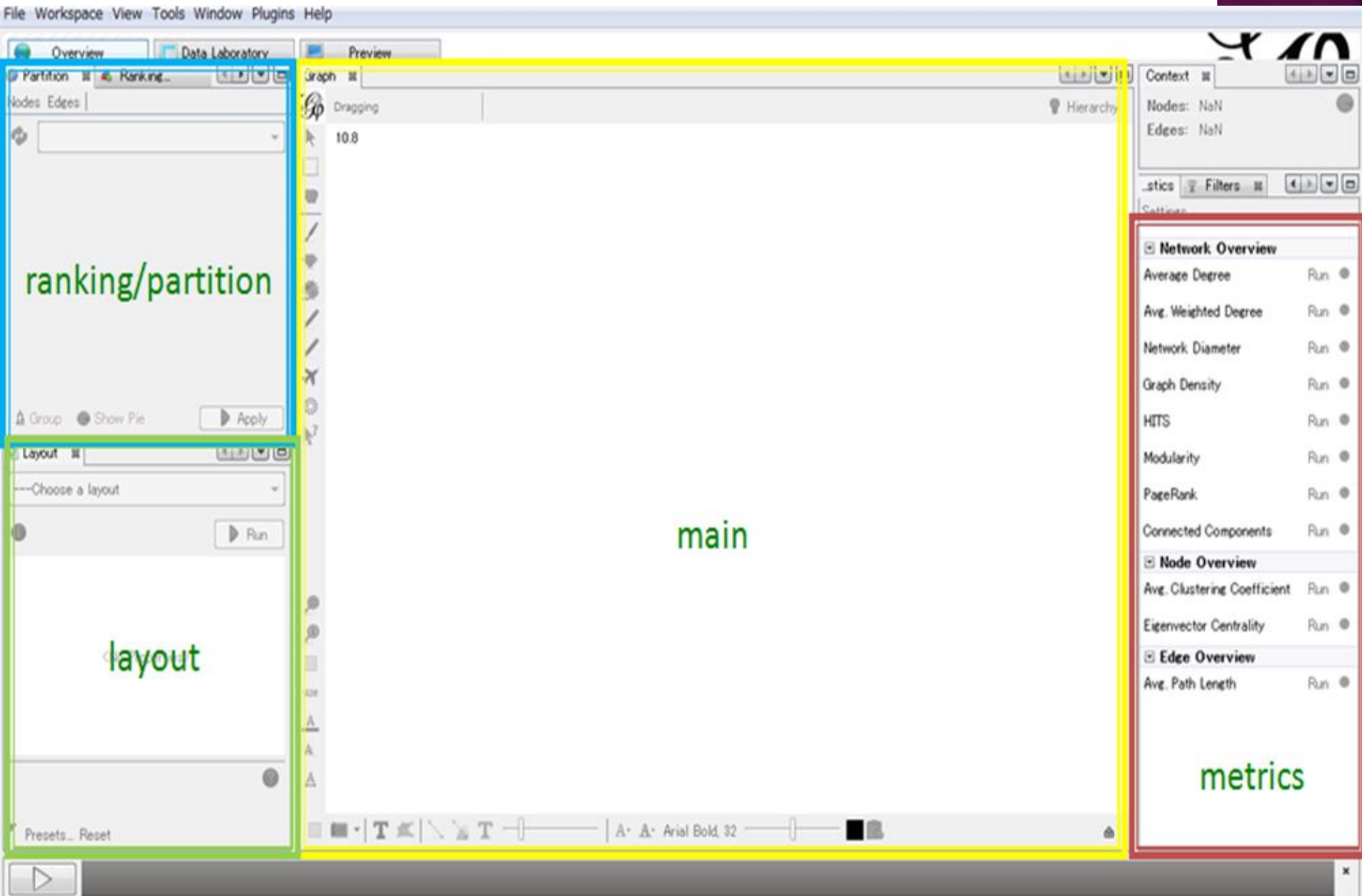
■ PNG

سه mode یا حالت زیر در منوی اصلی گفی در دسترس است که در mode امکانات خاص آن در صفحه نمایش داده می شود:

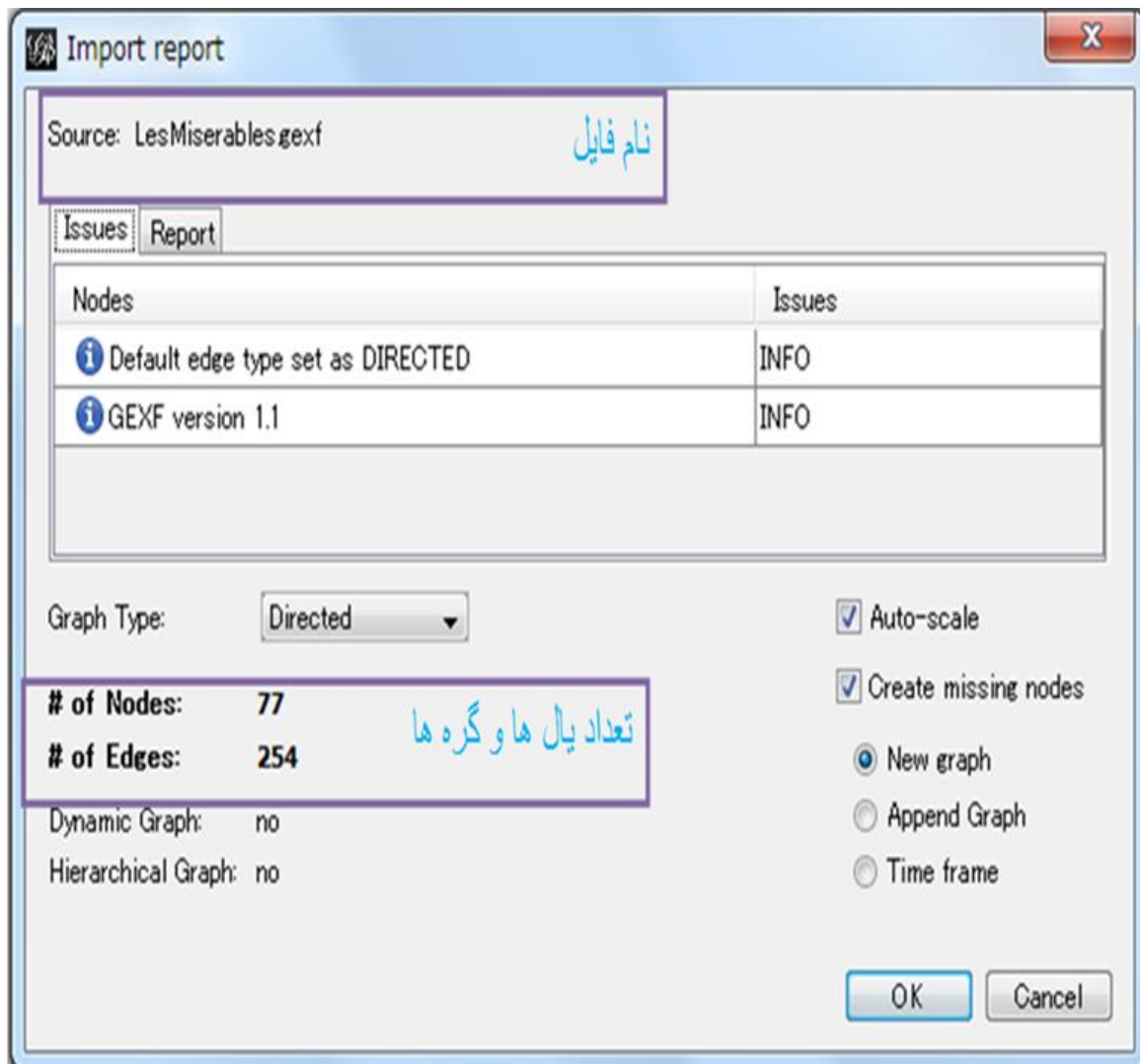
- **Overview:** graph manipulation mode.
- **Data Laboratory:** data visualization in tables.
- **Preview:** visual tuning before vectorial export.

○ در ادامه امکانات رابط گرافیکی گفی در مد Overview بررسی خواهد شد.

رابط کاربری نرم افزار گفی



منوی OPEN



○ پنجره : layout

○ پنجره Layout در گوشه پایین سمت چپ رابط کاربری گفی قرار دارد. LAYOUTها الگوریتم هایی هستند که گره ها را در فضای گرافیکی دو بعدی یا سه بعدی قرار می دهند. که در علم تحلیل گراف به آن بصری سازی یا بازنمایی می گویند.

○ پنجره معیارها یا (Metrics):

○ پنجره متریک ها در سمت راست صفحه قرار دارد و الگوریتم های آن شامل سه دسته زیر هستند:

○ دسته اول معیارهای شبکه : ۸ خانواده

○ دسته دوم معیارهای گره: ۲ خانواده

○ دسته سوم معیارهای یال: ۱ خانواده

○ معیارهای پویایی: ۴ خانواده

افزونه های گفی

 Plugins

Updates (8) Available Plugins (37) Downloaded Installed (74) Settings

Check for Updates

Search:

Update	Name	Category
☒	Project API	Gephi Core
☒	Filters Impl	Gephi Core
☒	Import API	Gephi Core
☒	Import Plugin UI	Gephi UI
☒	Processor Plugin	Plugin
☒	Export Plugin	Plugin
☒	Filters Plugin	Plugin
☒	Import Plugin	Plugin

Project API

Installed version: 0.9.2
Available version: 0.9.2.1
Author: Gephi Consortium
Date: 11/18/17
Source: Gephi Update Center
Homepage: <http://gephi.org/modules/project-api>

Plugin Description

API for projects and workspaces manipulation

Update

8 plugins selected, 11MB

Close

Help

جدول TABLE LAB

در این قسمت اطلاعات کامل گراف قابل نمایش است و امکان انجام تنظیمات و اعمال تغییرات بروی گراف وجود دارد. در جدول باز شده در قسمت بالا سمت چپ دو گزینه Nodes و Edges وجود دارد که اگر بروی هر کدام کلیک شود اطلاعات کامل تمام گره ها و یا یال های گراف را نمایش می دهد و یکی از مهمترین ویژگی های این قسمت قابلیت تعریف ستون در Data Table است که اگر بروی گزینه Formulic Merge کلیک شود یک صفحه باز می شود و امکان تعریف یک ستون جدید را می دهد که می توان مقدار این ستون را بر اساس پارامتر های عددی که بروی گراف محاسبه شده است تعیین کرد و در بخش های رتبه بندی یا رنگ بندی می توان از این ستون استفاده کرد .

جدول TABLE LAB

Overview Data Laboratory Preview

Data Table x

Nodes Edges Configuration Add node Add edge Formulaic Merge Search/Replace Import Spreadsheet Export table More actions Filter: Nodes

Nodes	Id	Label	Score	InDegr...	OutDegr...	TotalD...	PageRank	Stres...
moqawama.ir	-7106058891591636153	moqawama.ir	16.426432987843512	7.0	4.0	11.0	8.357903526617265E-4	308.0
alaska.blogfa.com	7286543556499367353	alaska.blogfa.com	16.426354410192452	0.0	0.0	0.0	3.2586042902690295E-4	0.0
turkdonyasi.loxblog.com	-847171131166398739	turkdonyasi.loxblog.com	16.426329579330172	0.0	1.0	1.0	3.2586042902690295E-4	0.0
ahestan.wordpress.com	8849237598880103295	ahestan.wordpress.com	16.426302754925267	18.0	27.0	45.0	0.0025830679052162592	31147.0
ahestan.ir	6355297051204749495	ahestan.ir	16.426291896801557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
maahinii.parsiblog.com	-8663238455874154149	maahinii.parsiblog.com	16.426264125501557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
jahanesanat.ir	-2102682939905156998	jahanesanat.ir	16.426252510801557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
bayaranali.persianblog.ir	-8096891004622328847	bayaranali.persianblog.ir	16.426248757201557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
akrane.persianblog.ir	-1114277791309929987	akrane.persianblog.ir	16.426229034801557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
changizi.blogfa.com	6546390385467949295	changizi.blogfa.com	16.426218783801557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
nasimesobh.persianblog.ir	8137100112420060418	nasimesobh.persianblog.ir	16.426190313001557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
mm220.persianblog.ir	1970896533499991363	mm220.persianblog.ir	16.426171895301557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
al-fateh.blogfa.com	620565031177082089	al-fateh.blogfa.com	16.426152017701557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
abadanjami.persianblog.ir	802526857021538500	abadanjami.persianblog.ir	16.426144122201557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
fadaeiagha.parsiblog.com	8360553921991672209	fadaeiagha.parsiblog.com	16.426124309101557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
rezaeibar.blogfa.com	2488648336356405681	rezaeibar.blogfa.com	16.426123115701557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
1001shayadobayad.mihanblog.com	6336602246475345703	1001shayadobayad.mihanblog.com	16.426078733601557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
amirihosain.blogfa.com	6243224573297890942	amirihosain.blogfa.com	16.426068697301557	0.0	0.0	0.0	3.2586042902690295E-4	0.0
farsiblog.mihanblog.com	1027436036580870954	farsiblog.mihanblog.com	16.426068157404703	2.0	8.0	10.0	4.8645742388584974E-4	3463.0
shere-man2.blogfa.com	8744663346947350989	shere-man2.blogfa.com	16.426066517088906	0.0	0.0	0.0	3.2586042902690295E-4	0.0
namin1387.persianblog.ir	-5383215334287659615	namin1387.persianblog.ir	16.426061204941377	2.0	1.0	3.0	5.784483781804033E-4	0.0
jalabnews.loxblog.com	5619544461890159920	jalabnews.loxblog.com	16.4259976883523	0.0	0.0	0.0	3.2586042902690295E-4	0.0
rahpooyesh.blogfa.com	6038854542250032765	rahpooyesh.blogfa.com	16.42599641918473	3.0	11.0	14.0	4.930544130783983E-4	941.0

Formulaic Merge

Score
InDegree
OutDegree
TotalDegree(directed)
PageRank
Stress Centrality

Nodes Attributes Edges Attributes

OK Cancel

Add column Merge columns Delete column Clear column Copy data to other column Fill column with a value Duplicate column Create a boolean column from regex match Create column with list of regex matching groups Negate boolean values Convert column to dynamic

◉ با استفاده از پنجره **Filters** می‌توان بروی ویژگی‌هایی از گراف اعمال محدودیت کرد. این ماژول از چند قسمت اصلی تشکیل شده است که در ادامه به آنها اشاره می‌کنیم.

Attributes

۱- Equal

در این قسمت می‌توان دقیقا یک نود یا یک یال یا یک خوشه و یا هر معیار دیگری که روی گراف محاسبه شده است را فیلتر کرد.

۲- Partition

در این قسمت می‌توان بروی خوشه‌های محاسبه شده بروی گراف اعمال محدودیت کرد مثلا یک یا چند خوشه مورد نظر را انتخاب و گراف جدید را بر اساس این انتخاب‌ها تشکیل داد.

فیلترینگ

۳- Partition Count

در این قسمت می توان محدوده ایی از تعداد خوشه های موجود را محدود کرد .

۴- Range

در این قسمت می توان بروی معیارهایی که نتیجه آنها عدد هستند محدودیت مورد نظر را اعمال کرد .

Edges

Edge weight

در این قسمت می توان تعداد خاصی از یال ها را فیلتر کرد .

فیلترینگ

Operator

۱- Not(Edges)

در این قسمت می توان تمام یال ها را بدون حضور نود ها نمایش داد .

۲- Not (Nodes)

در این قسمت می توان تمام نود ها را بدون حضور یال ها نمایش داد .

گراف کاوی

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:
علیرضا حبیبی

گراف کاوی

- تشخیص انجمن
- تشخیص مرکزیت
- بصری سازی
- پیشبینی لینک
- نمونه برداری
- تحلیل انتشار

فهرست مطالب

◉ انواع گراف تصادفی

◉ انواع مرکزیت

■ Power

◉ انواع تشخیص انجمن

■ Louvain

◉ انواع پیشبینی ارتباط

■ Adamic Adar

◉ انواع بصری سازی

■ رنگ آمیزی

■ وزن دهی

○ Degree

○ inDgree

■ بازنمایی

○ ForceAtlas

○ Fruchterman Reingold

○ Yifan Hu

○ OpenOrd

◉ انواع نمونه برداری

■ ۱

بصری سازی

www.bigdata-ir.com

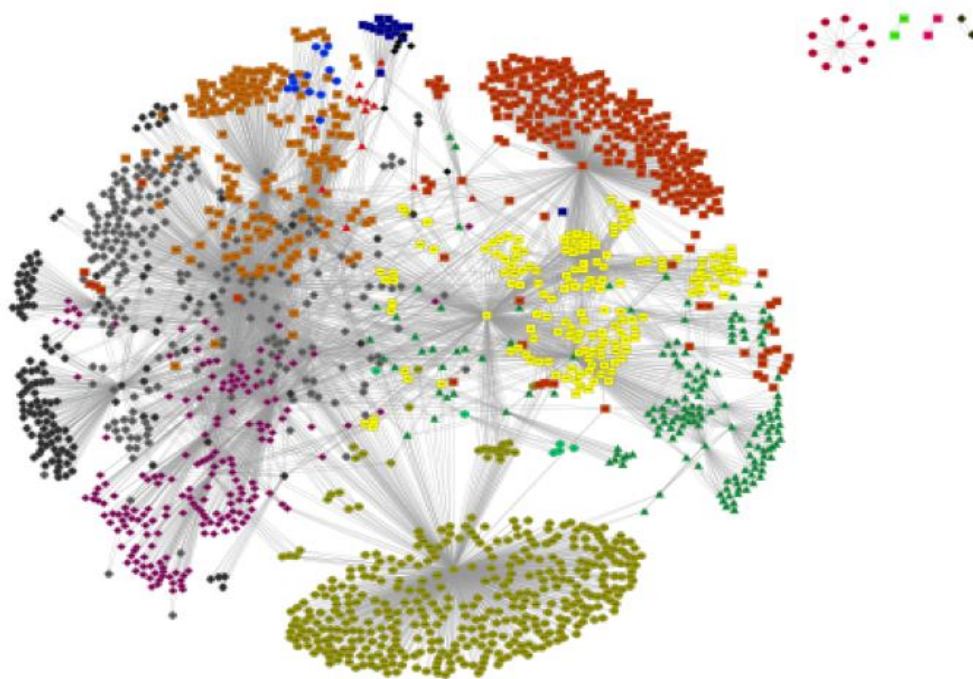
Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

الگوریتم های بازنمایی

برای نظم دادن به نحوه نمایش گراف می توان از طریق پنجره Layout که در حالت معمول در پایین سمت چپ صفحه گفی قرار دارد اقدام نمود. الگوریتم های مختلفی برای انتخاب آرایش گراف وجود دارند که هر یک بر اساس روشی گراف را مرتب می کنند که در ادامه به توضیح آنها می پردازیم .



رویکردهای بصری سازی مبتنی بر اعمال نیرو

⊙ نکته مهم: بصری سازی با تشخیص انجمن متفاوت است

⊙ این رویکردها با توجه به دو عامل زیر یکی از دو نیروی جاذبه و دافعه را

بین نودها اعمال می کنند:

■ آیا یالی بین دو نود وجود دارد؟

■ آیا موقعیت نودها (فاصله مابین آنها) از هم دور است؟

نیروی جاذبه

○ نیروی جاذبه در این رویکردها مشابه نیروی جاذبه فنر که بر اساس قانون

هوک (Hooke) محاسبه می شود:

$$F=kx$$

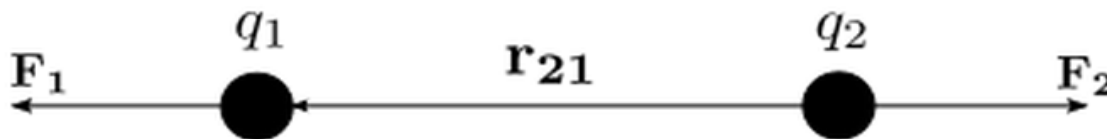
■ F برابر با نیرو، و x نشان دهنده فاصله می باشد.

■ k یک ثابت است .

نیروی دافعه

○ نیروی دافعه در این رویکردها مشابه شارژ الکتریکی می باشد. به این صورت که نودهایی که به یکدیگر یالی ندارند بصورت ذراتی می باشند که به یکدیگر نیروی دافعه وارد می کنند.

○ نیروی دافعه شارژ الکتریکی با استفاده از قانون کولمب (Coulomb) محاسبه می شود.



- ◉ ForceAtlas (FA)
- ◉ Fruchterman Reingold (FR)
- ◉ Yifan Hu (YH)
- ◉ OpenOrd (OO)

◉ این الگوریتم ها در دسته Transformation قرار دارند و فقط گراف در جهت خاص یا در یک مقیاس کوچک یا بزرگ می کنند.

- “Clockwise Rotate” with angle -90°
- “Counter-Clockwise Rotate” with angle 45°
- “Expansion” with scale factor 1.2
- “Contraction” with scale factor 0.8

جلوگیری از هم پوشانی گره ها

◉ به منظور جلوگیری از هم پوشانی گره ها چند روش وجود دارد

۱- استفاده از الگوریتم Node OverLapping

این الگوریتم را معمولا بعد از الگوریتم بازنمایی دیگر اجرا می کنیم تا هم پوشانی های موجود را از بین ببرد و از مهمترین پارامترهای آن در ادامه ذکر شده است .

Speed

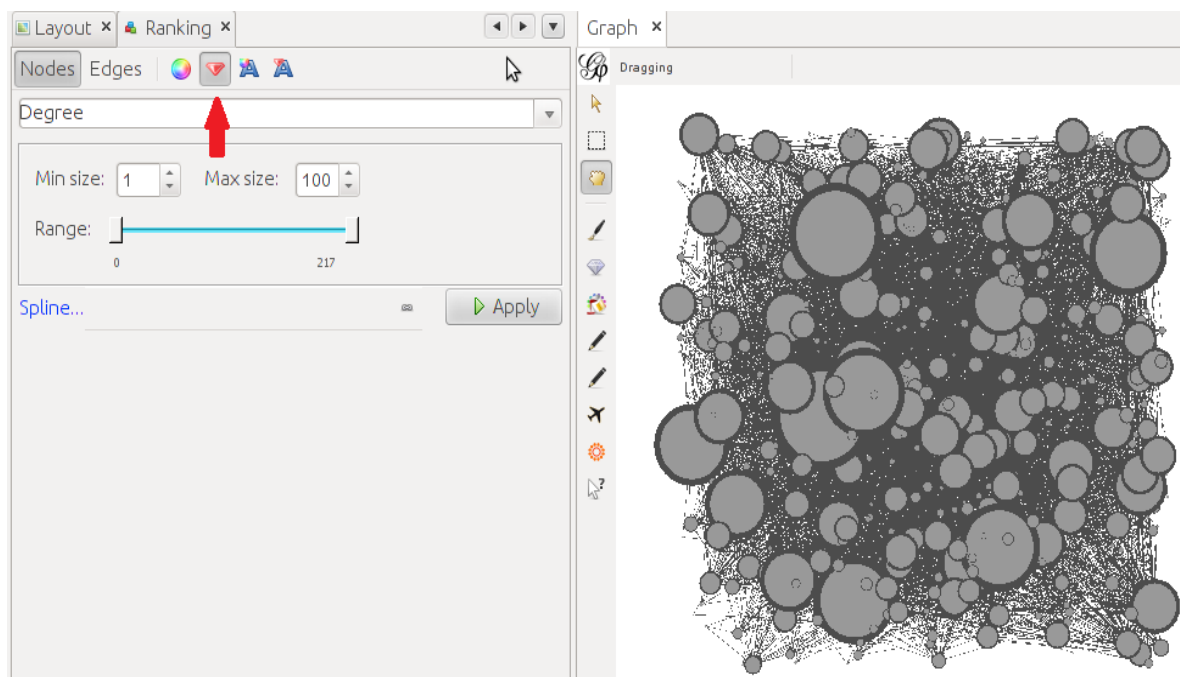
سرعت الگوریتم را تعیین می کند و اگر مدت زمان اجرا اهمیت چندانی ندارد مقدار آن را یک قرار بدهیم عملکرد بهتری دارد .

Margin

فاصله بین گره را تعیین می کند

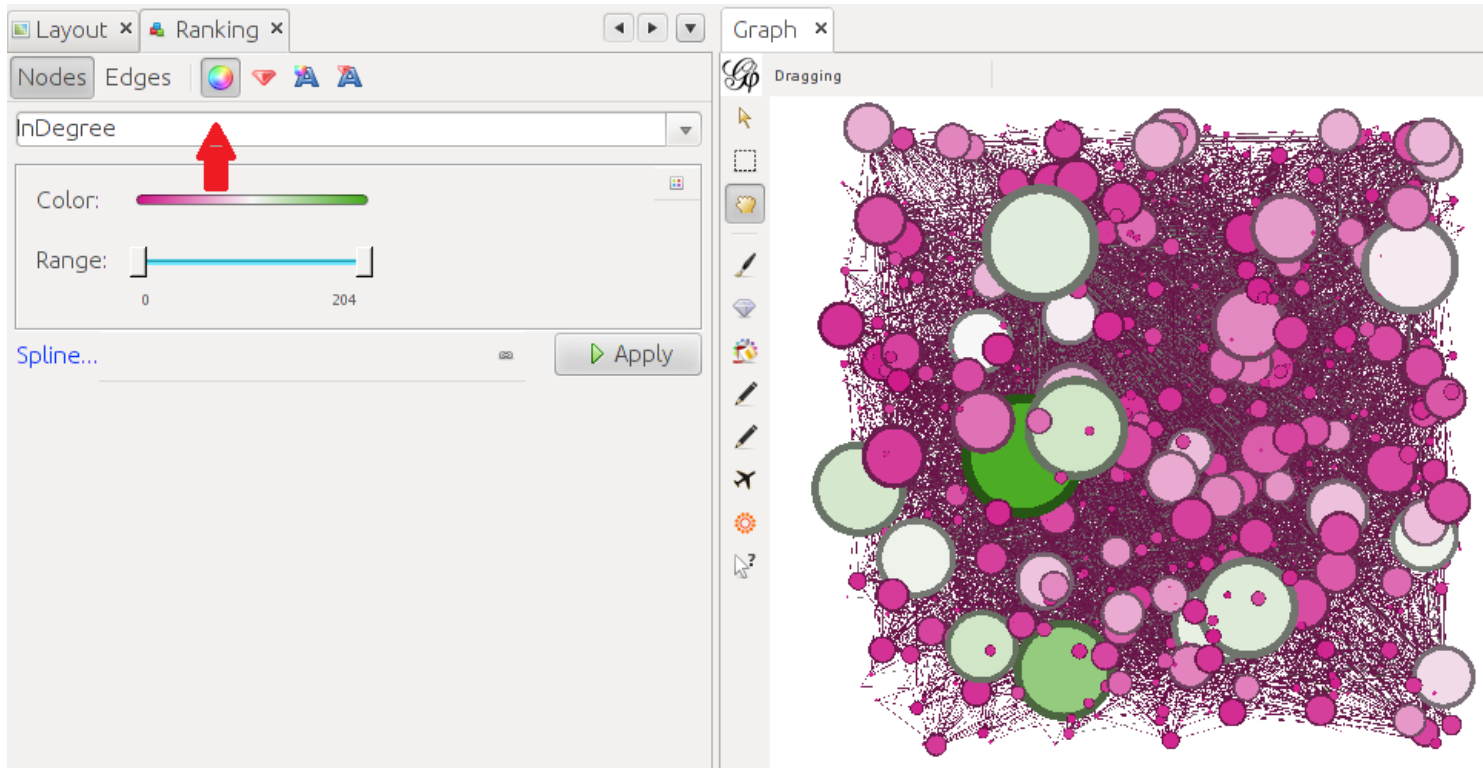
رتبه بندی

یکی از پرکاربردترین ماژول های این سامانه ،ماژول رتبه بندی می باشد . در این قسمت می توانیم رنگ و اندازه گره ، رنگ و اندازه برچسب ها (label) را بر اساس معیارهای عددی که بروی گراف محاسبه شده اند تغییر بدهیم .به طور مثال اندازه گره ها را بر اساس معیار Degree محاسبه می کنیم .



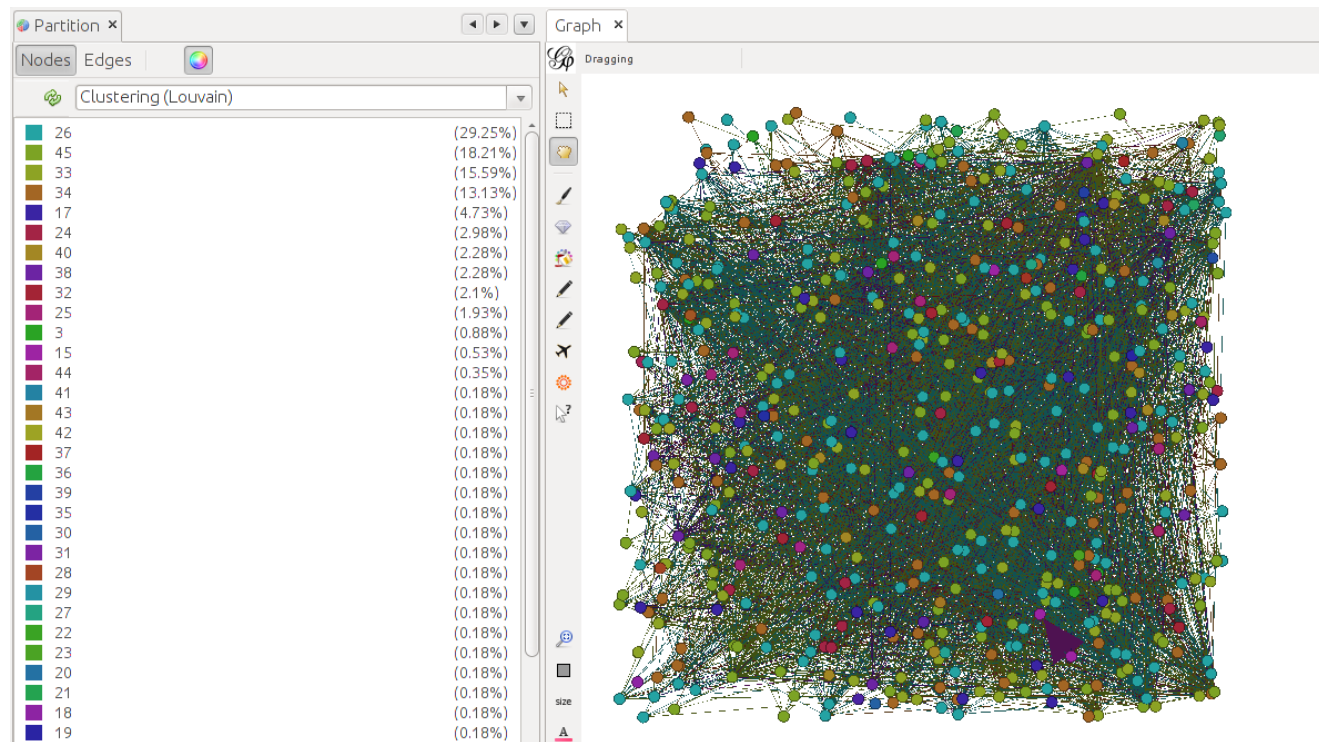
رتبه بندی

در این قسمت ما رنگ گره ها را بر اساس InDegree محاسبه می کنیم .



رنگ بندی

در این قسمت ما می توانیم بر اساس معیار های عددی صحیح محاسبه شده بروی گراف به گره های موجود در گراف رنگ نسبت بدیم تا با متمایز شدن از یکدیگر دید بهتری از گراف پیدا کنیم .



مرکزیت

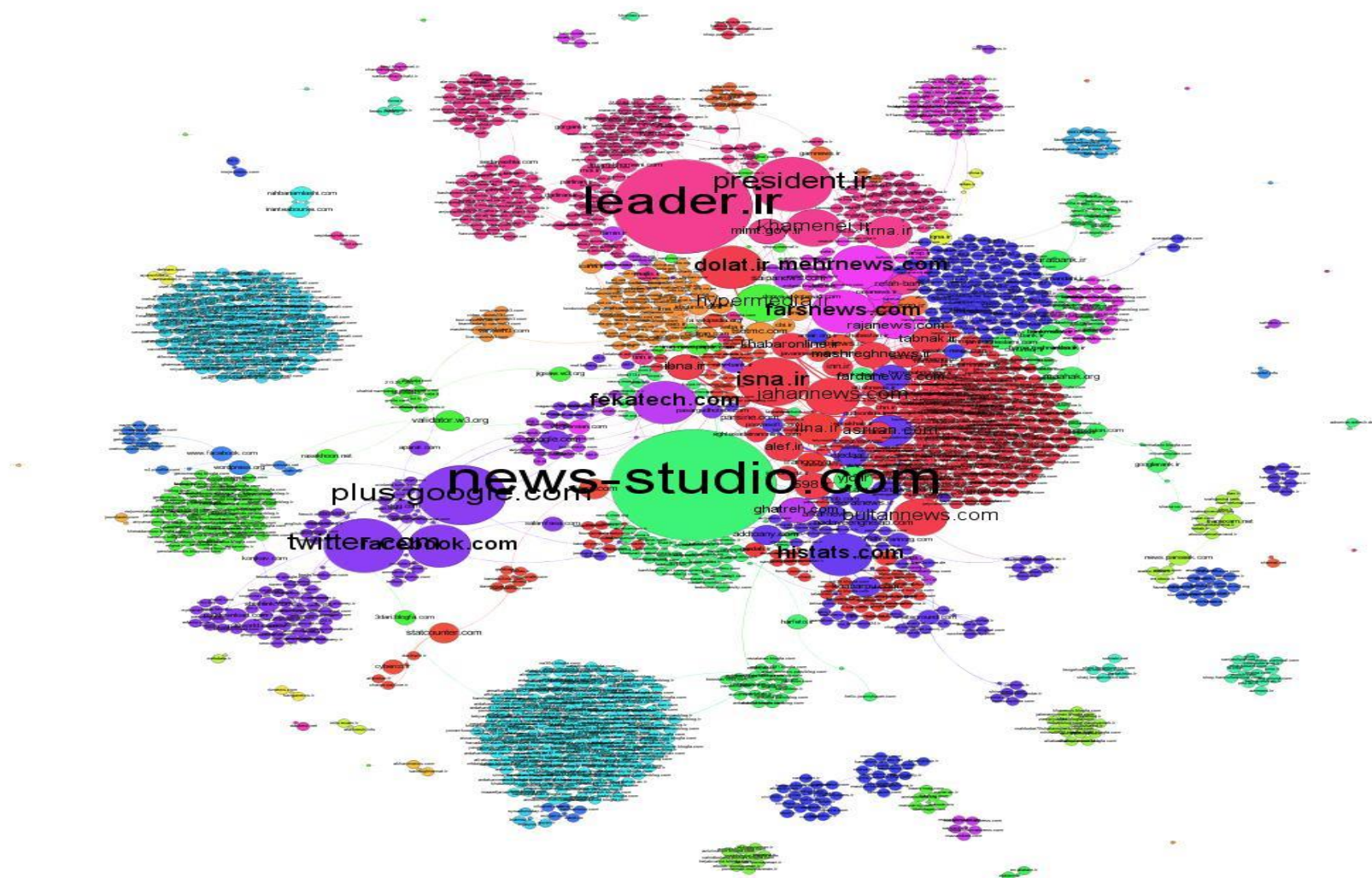
www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

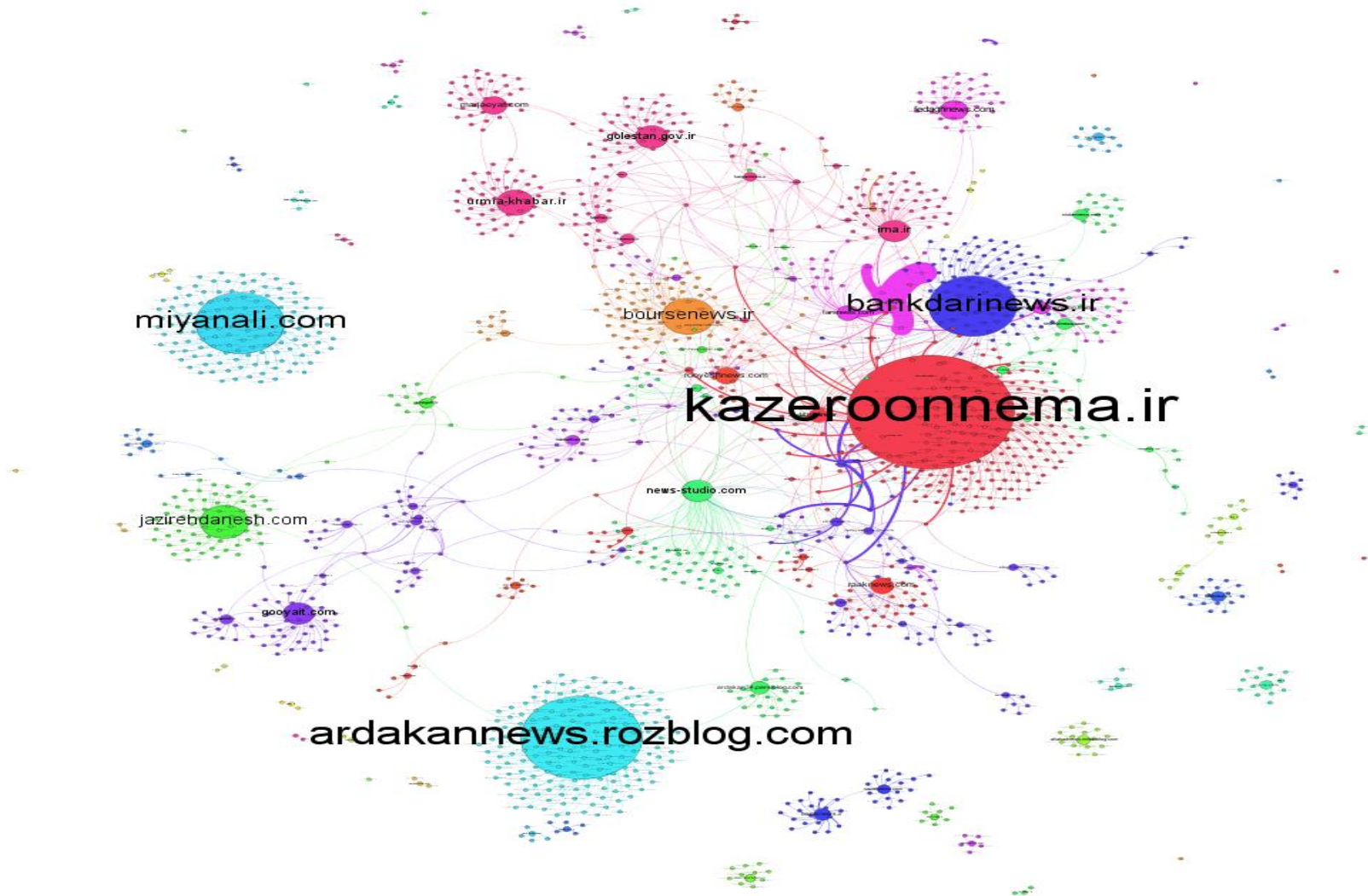
مدرس:

علیرضا حبیبی

درجه ورودی



درجه خروجی

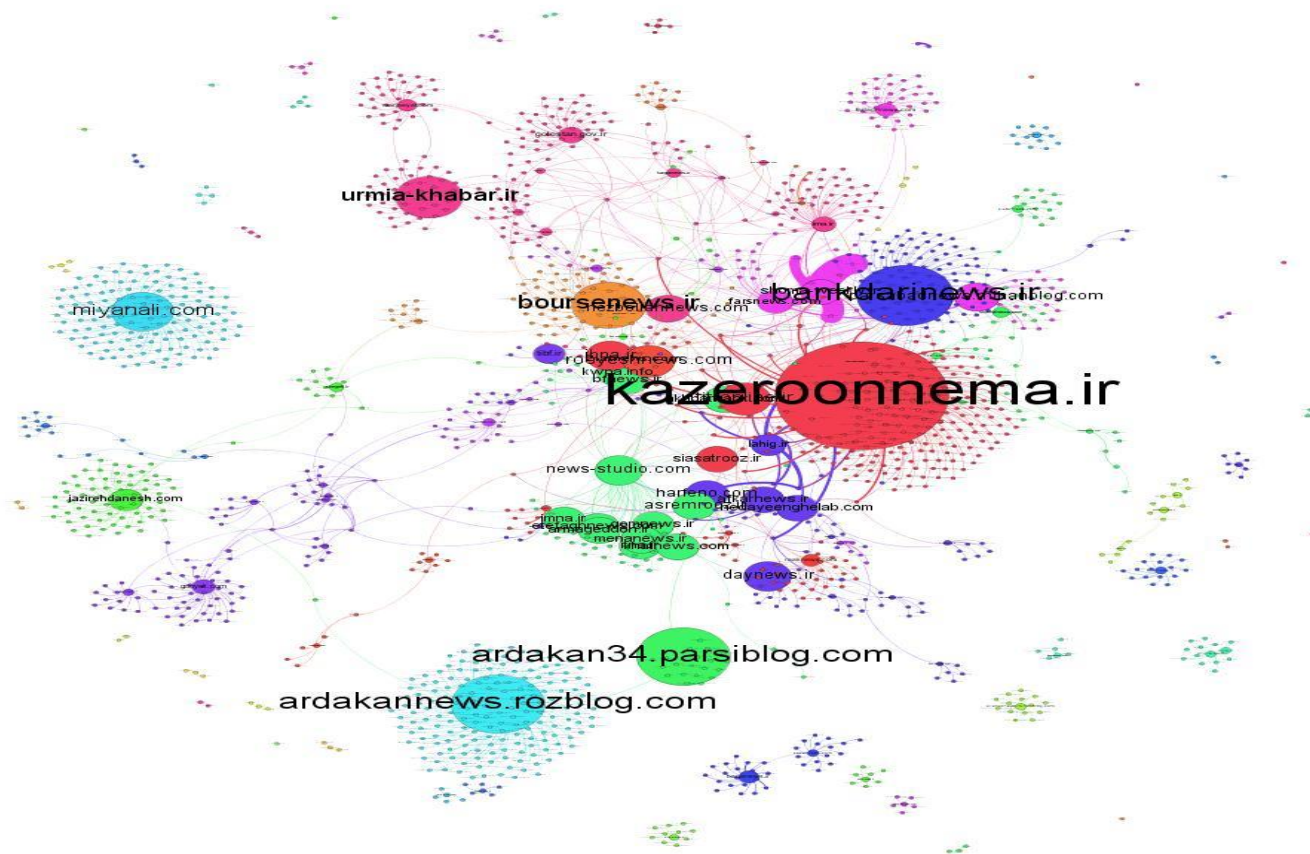


بینابینی BETWEENNESS



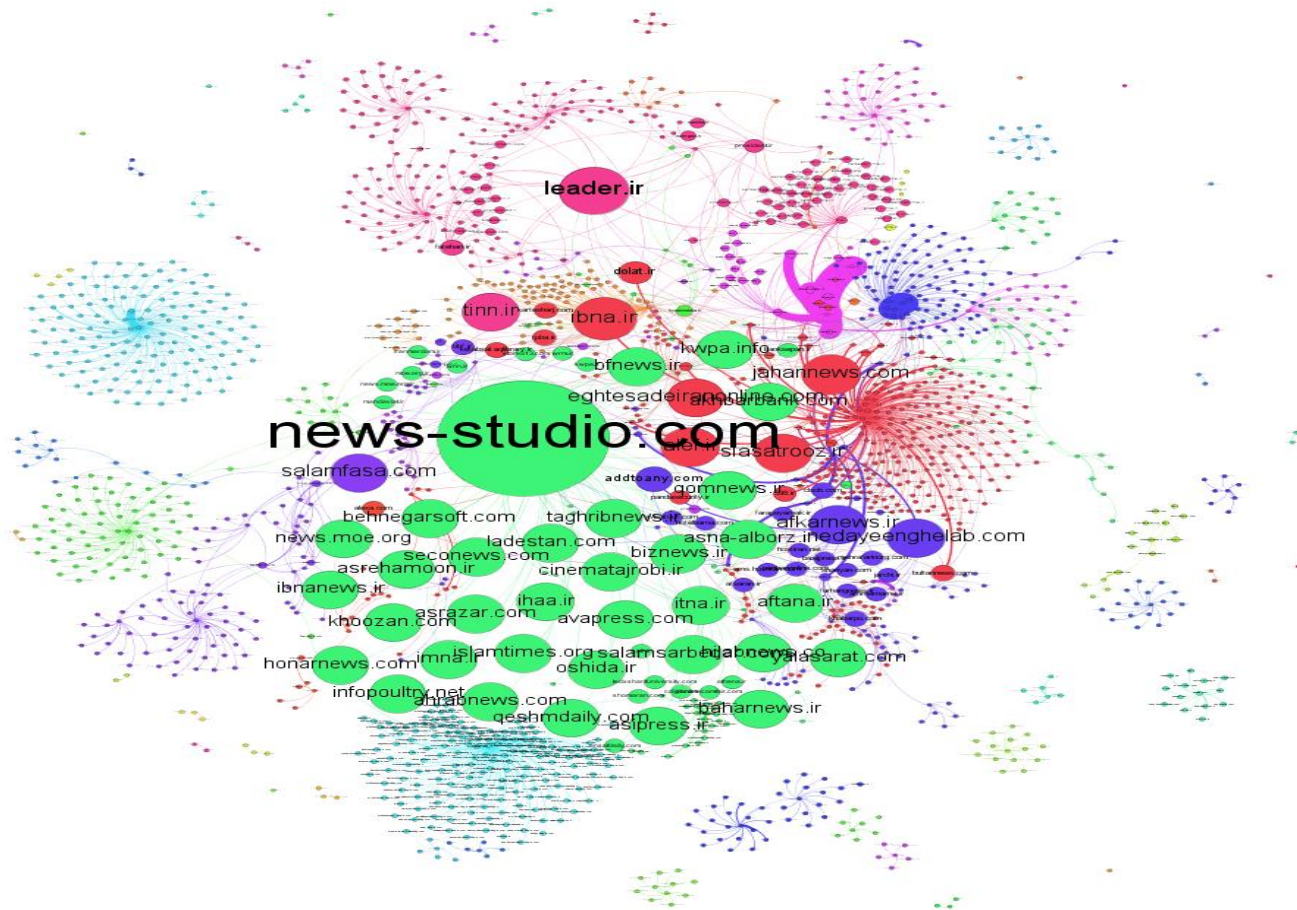
دسترسی گره ها

REACH



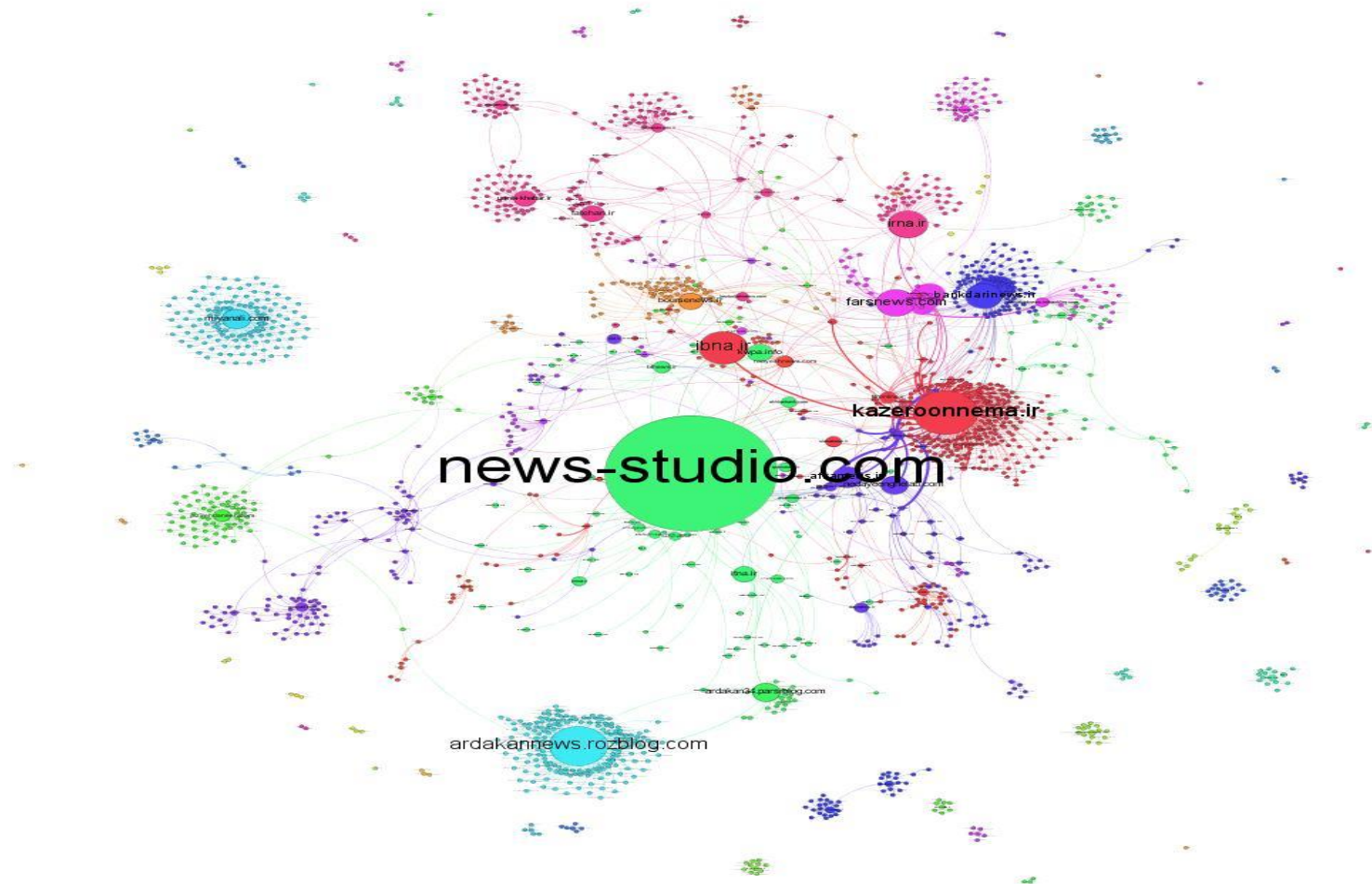
$$R(3)_{(kazeroonnema.ir)}=0.2$$

بردار ویژه EIGEN VECTOR



آسیب پذیری گره

NODE VULNERABILITY



معیارهای مرکزیت

توپولوژی گراف

- فراوانی درجات گره ها (Degree distribute)
- ضریب کلوپ پولداران (Rich club)
- قطر گراف (Diameter)
- ضریب تراگذری (Transitivity coefficient)
- ضریب خوشه بندی (Clustering Coefficient)
- محاسبه کوتاهترین مسیر میان گرههای گراف
- فاصله متوسط گرهها از هم
- بینابینی (Betweenness)
- نزدیکی (Closeness)
- دوری از مرکز (Eccentricity)

معیارهای مرکزیت

- دسترسی گره ها (Reach)
- کارایی گره (Efficiency)
- تنش (Stress)
- قدرت (Power)
- بردار ویژه (Eigen Vector)
- آسیب پذیری گره (Node vulnerability)
- رتبه صفحه (Page rank)
- Hits

ضرب کلوپ پولداران

RICH CLUB

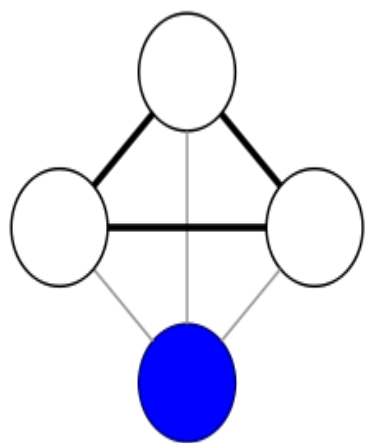
- این معیار نشان دهنده میزان تمایل گره های با درجات بالا (که قشر مهمتر شبکه می باشند) به همکاری با یکدیگر می باشد.
- این مقدار می تواند شاخصی برای تشخیص آستانه تحمل یک شبکه در مقابل از کار افتادن گره های مهم باشد.

$$\text{تعداد یالهایی بین گره های با درجه بزرگتر از } k = \frac{\text{تعداد یالهای کل ممکن بین آنها}}{k}$$

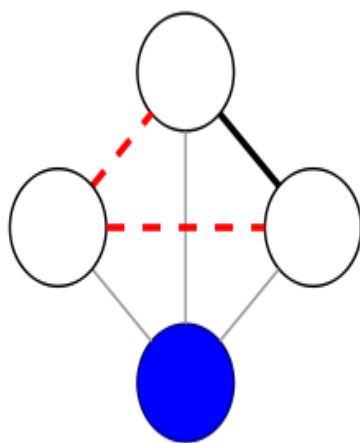
ضریب خوشه بندی

CLUSTERING COEFFICIENT

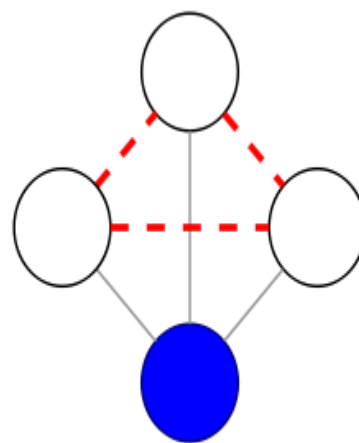
- ضریب خوشه بندی، احتمال اتصال گره های همسایه یک گره به یکدیگر است.
- نسبت تعداد یال های موجود بین گره های مجاور یک گره به تعداد یال های که ممکن است بین آنها وجود داشته باشد.
- این معیار برای کل گراف، میانگین ضرایب خوشه بندی همه گره ها می باشد.
- هر چه این احتمال بیش تر باشد، یعنی اینکه گره ها تمایل بیشتری برای تشکیل گروه دارند.



$$C = 1$$



$$C = 1/3$$



$$C = 0$$

ضریب تراگذری

TRANSITIVITY COEFFICIENT

- این معیار همانند ضریب خوشه بندی نشان دهنده میزان تمایل گره های یک گراف به همکاری با یکدیگر است ولی بر خلاف ضریب خوشه بندی که برای تک تک گره هاست، این معیار به بررسی میزان بر روی کل گراف می کند.
- نشان می دهد اگر فرد A با B دوست باشد و B با C چقدر احتمال دارد A نیز با C دوست باشد (برای گرافهای جهت دار).
- نسبت مثلثهایی که در کل گراف وجود دارد به تعداد کل مثلثهایی که می توانند وجود داشته باشند.

بینابینی

BETWEENNESS

بینابینی عبارت است از نسبت تعداد دفعاتی که یک گره یا یک یال بر روی کوتاهترین مسیر میان نودهای مختلف یک گراف قرار می گیرد.

کاربرد:

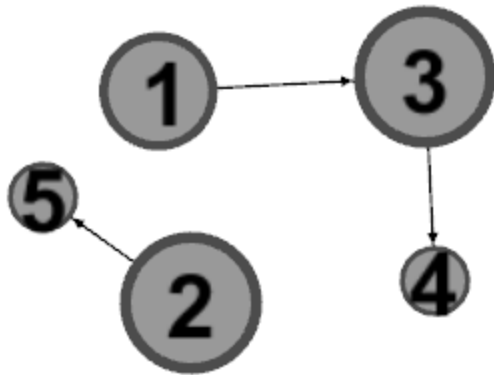
- محل افرادی که توانایی مرتبط ساختن با گروه های دیگر را دارند نشان می دهد.
- نشان دهنده درصدی از اطلاعات است که از یک گره می گذرد.
- مشخص کننده توانایی یک گره برای تسهیل گسترش ارتباط بین سایر عناصر گره های گراف است و در واقع نمایشی برای میزان قابلیت هر گره برای کمک به دسترسی سایرین به اطلاعات و یا گسترش یک تاثیر در شبکه می باشد.

بینایی BETWEENNESS



نزدیکی CLOSENESS

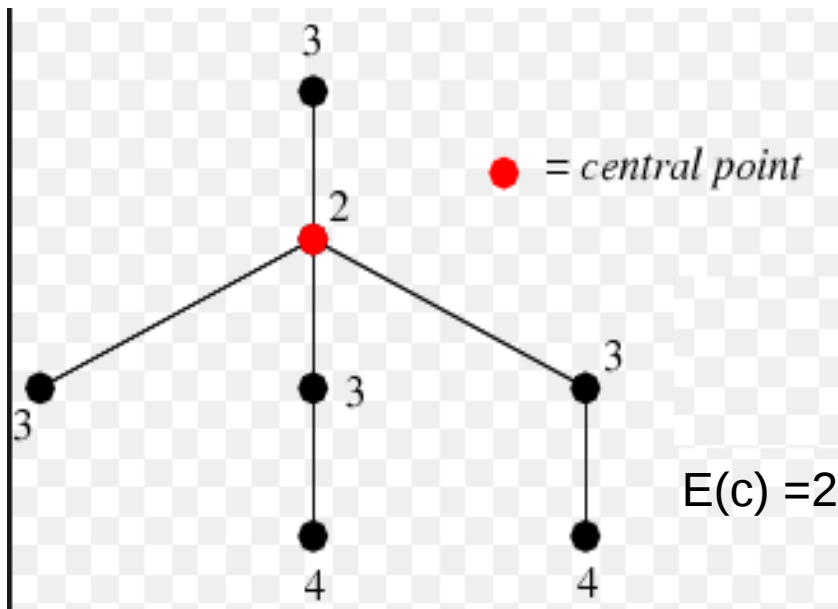
- نزدیکی گره X برابر است با عکس متوسط کوتاهترین فاصله گره X تا سایر گره های گراف.
- به عبارتی وسط ترین نقطه گراف از جهت نزدیک بودن به بقیه نودها
- چه مدت زمان می برد که اطلاعات از یک گره به دیگر گره ها برسد (گره هایی که به آن دسترسی دارد).



۱	۲	۳	۴	۵
۰.۶	۱	۱	۰	۰

دوری از مرکز ECCENTRICITY

- پیدا کردن طولانی ترین مسیر از میان کوتاه ترین مسیرهایی است که به گره های دیگر دارد.
- این معیار می تواند میزان دوری فاصله یک گره به گره ها دیگر را بسنجد.
- برعکس معیار نزدیکی



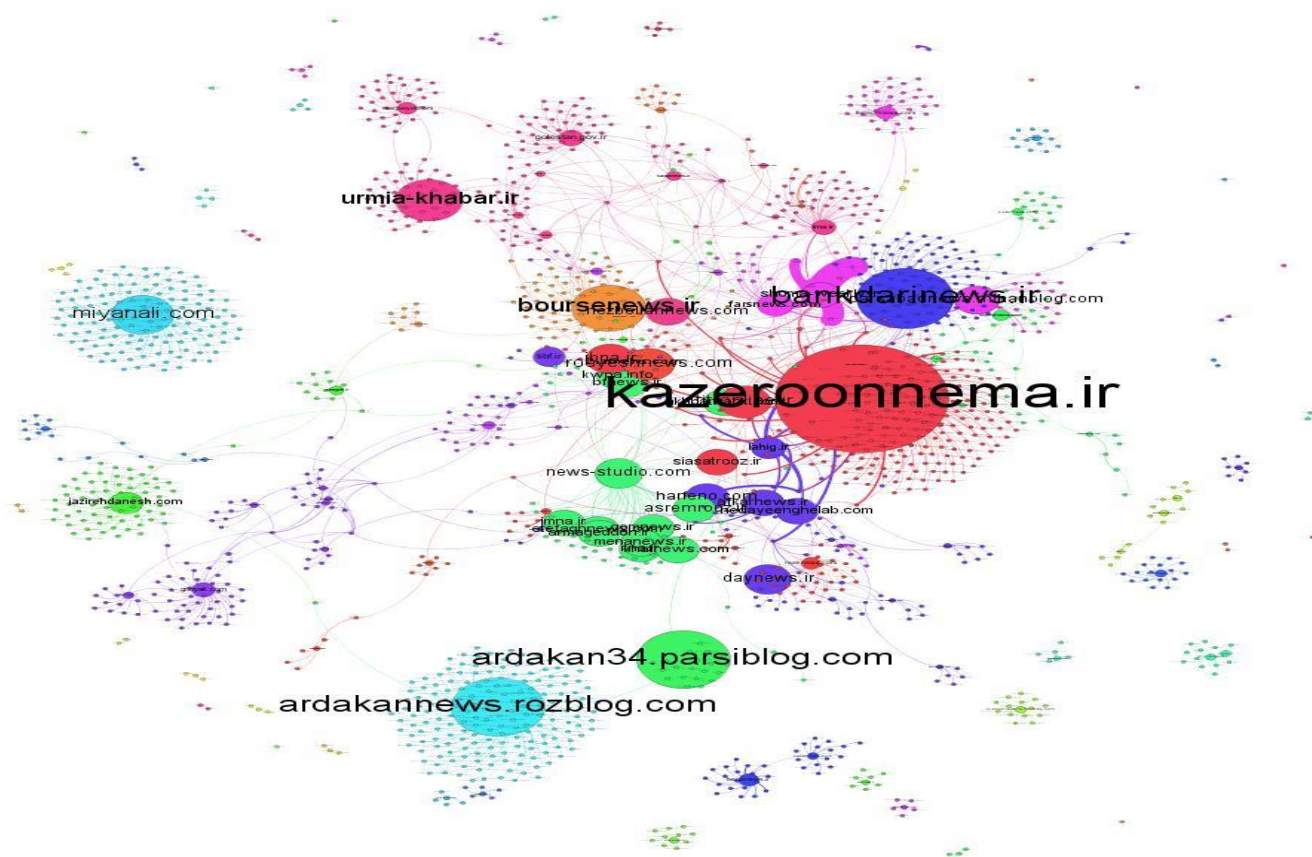
دسترسی گره ها

REACH

- **REACH:** تعداد گره هایی است که می توانند در زمانی کمتر از k قدم، از خبری که توسط گره مبدا منتشر شده است، مطلع شوند.
- درصد تعداد گره هایی که از هر گره با حداکثر k قدم در دسترس هستند.
- در حالتی که $k=1$ مقدار این معیار مشابه با مرکزیت درجه خواهد بود.

دسترسی گره ها

REACH



$$R(3)_{(kazeroonnema.ir)}=0.2$$

کارایی گره

EFFICIENCY

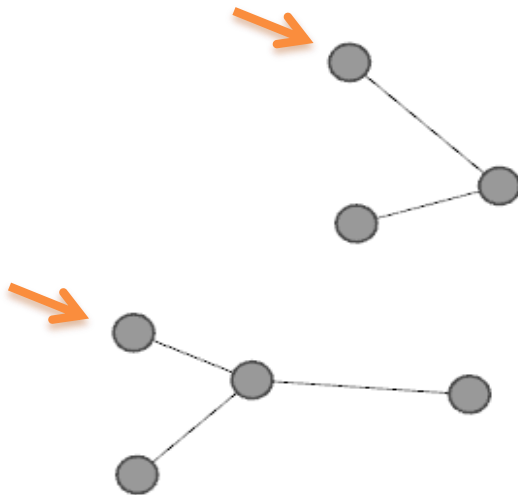
- معکوس متوسط کوتاهترین مسیر های میان هر دو گره
- این معیار به این معناست که یک شبکه چقدر در مبادله اطلاعات بین گره هایش خوب عمل می کند.
- هرچه مسیر بین گره ها کوتاهتر باشد، مسیر برای انتقال اطلاعات کاراتر خواهد بود.
- برای محاسبه کارایی گراف، از کارایی تمام مسیرهای موجود بین هر دو گره از گراف میانگین گرفته می شود.
- مثلاً کارایی یک گراف بدون جهت بیشتر از گراف جهت دار است. چرا؟

تنش STRESS

- گره ای دارای تنش و حساسیت بالاست که تعداد کوتاه ترین مسیرهای زیادی از آن می‌گذرد.
- این معیار نشان‌دهنده تعداد مسیرهای کوتاه، بین هر دو گره دیگر است که از یک گره خاص می‌گذرد.
- این معیار مشابه معیار مرکزیت بینابینی است با این تفاوت که بینابینی نشان‌دهنده درصدی از اطلاعات کل گراف است که از یک گره می‌گذرد، حال آنکه در معیار تنش، وجود مسیرهای زیاد در یک جز از گراف برای گسترش اطلاعات، بر حساس بودن آن کفایت می‌کند.

قدرت POWER

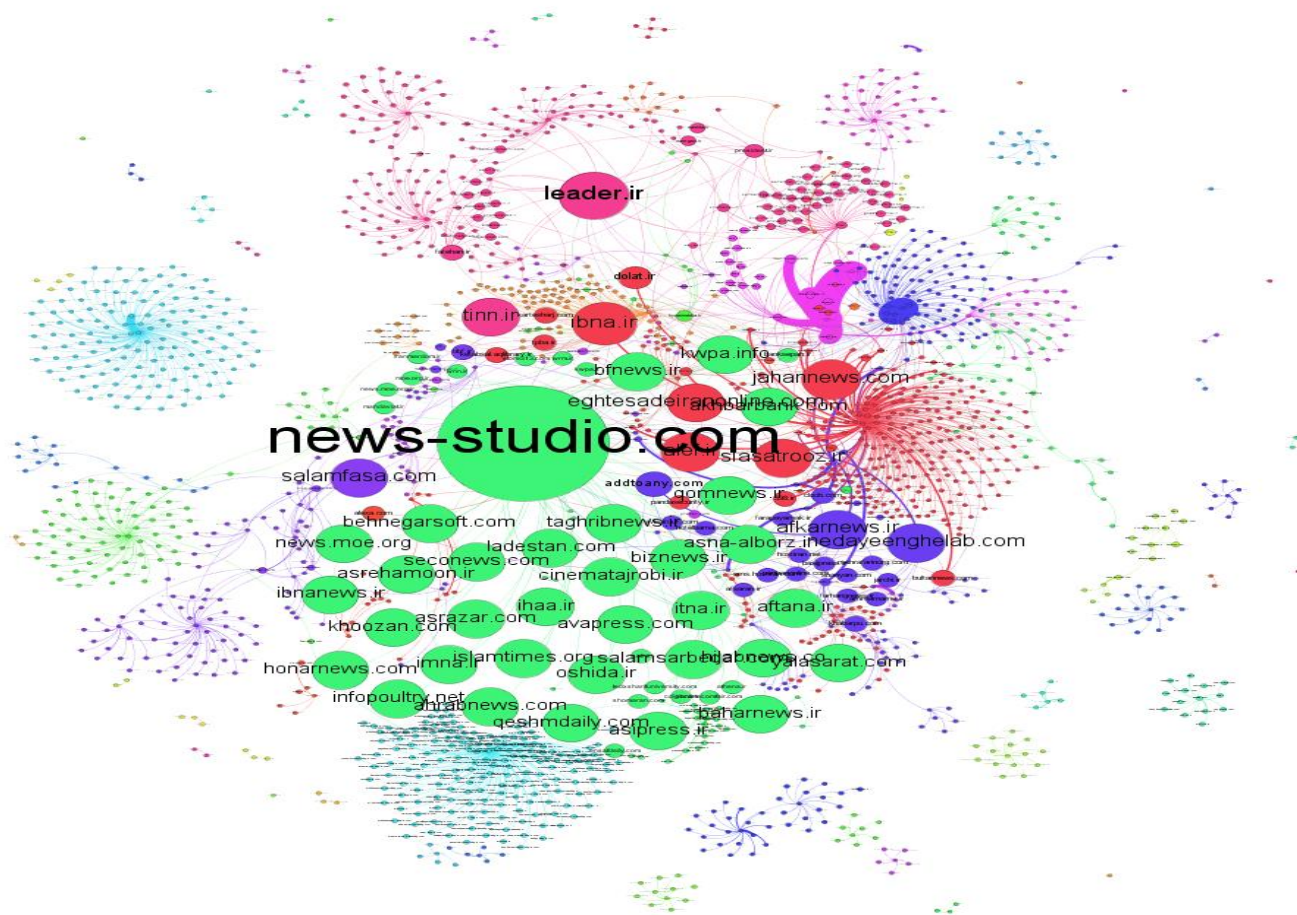
- این معیار مشابه معیار مرکزیت درجه می‌باشد. با این تفاوت که برای تخمین مرکزیت یک گره، علاوه بر استفاده از تعداد همسایه‌های محلی آن گره (گره‌هایی که مستقیماً با آن گره در ارتباط هستند)، درصدی از تعداد گره‌هایی را که آن گره در محدوده بزرگتری با آنها در ارتباط غیر مستقیم است را در نظر می‌گیرد
- و قدرت یک گره را بر اساس قدرت همسایگانش تخمین می‌زند.



بردار ویژه EIGEN VECTOR

- مرکزیت بردار ویژه هر گره معیاری برای تشخیص اهمیت آن گره در یک شبکه می باشد و در واقع تعمیمی از مرکزیت درجه است.
- گرهی که با تعداد n گره مهم در ارتباط می باشد، به نسبت گرهی که با n گره غیر مهم در ارتباط می باشد، جایگاه ویژه تری در شبکه دارد.

بردار ویژه EIGEN VECTOR

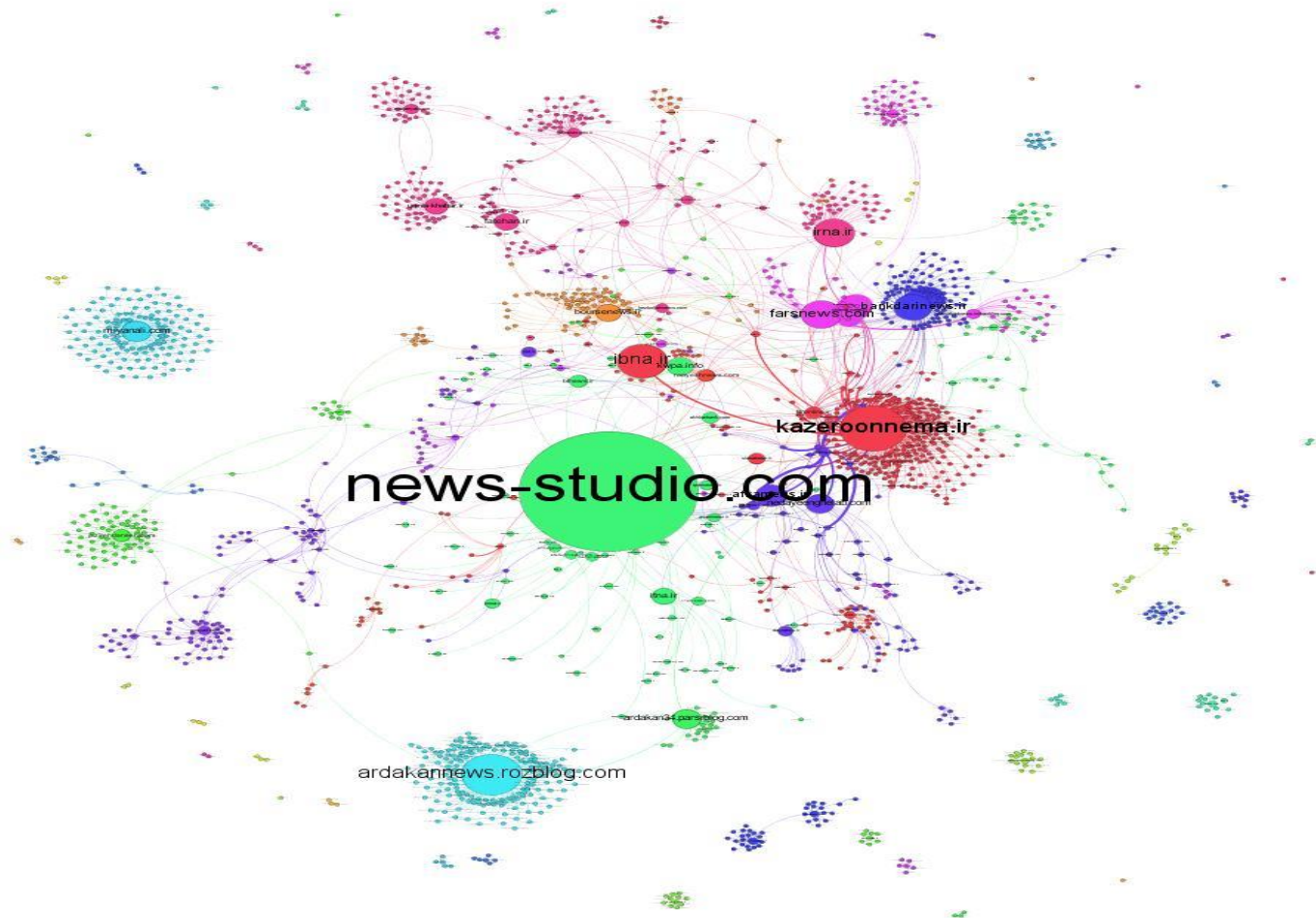


آسیب پذیری گره

NODE VULNERABILITY

- این معیار برای تک تک گره ها محاسبه می شود و بر اساس آن می توان میزان اهمیت هر گره در شبکه را مشخص نمود.
- گره هایی که آسیب پذیری بالاتری دارند، شبکه نسبت به حذف آنها حساس تر بوده و نقش آنها در شبکه گراف مهم تر می باشد.
- می توان تشخیص داد که کدام گره ها از اهمیت بیشتری در شبکه برخوردار می باشند و می بایست نسبت به مصون کردن این گره ها در شبکه تدابیری اندیشید.
- برای محاسبه این معیار هر بار یک گره را حذف می کنیم و کارایی شبکه را در نبود آن گره محاسبه می نماییم تا مشخص شود شبکه چه میزان کارایی اولیه خود را از دست داده است.

آسیب پذیری گره NODE VULNERABILITY



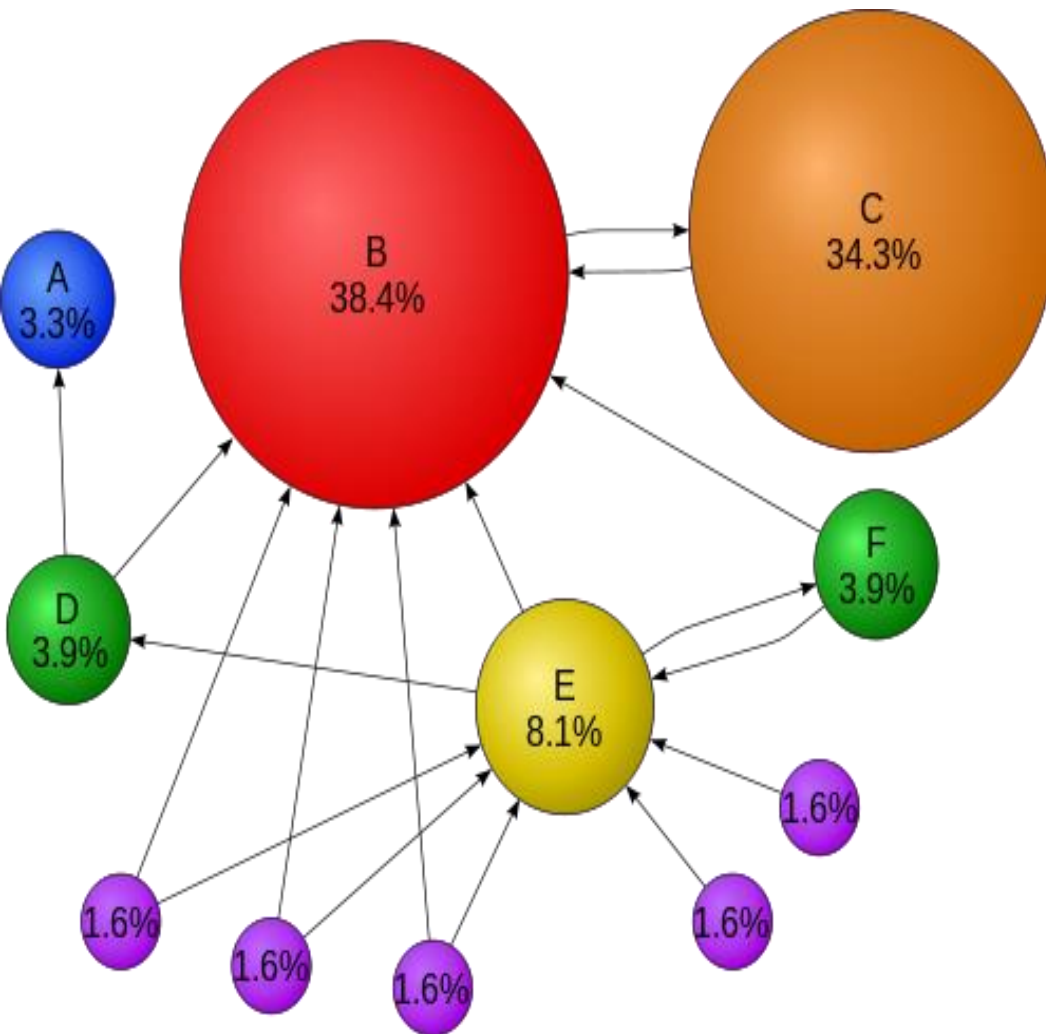
رتبه صفحه

PAGE RANK

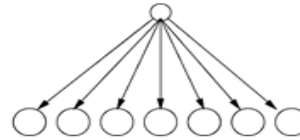
- این الگوریتم برای رتبه‌دهی به صفحات وب و تعیین ارزش آنها بر اساس پیوندهای بین آنها می‌باشد.
- در این الگوریتم، میزان ارزش و اعتبار یک صفحه به ارزش و اعتبار صفحاتی وابسته است که به او لینک داده‌اند و هرچه این صفحات با ارزش‌تر باشند اعتبار صفحه بالاتر می‌رود.
- با توجه به اینکه در ابتدا دانشی در مورد ارزش صفحات نداریم در این الگوریتم ابتدا به تمام صفحات مقدار یکسانی به عنوان ارزش اولیه داده می‌شود.
- چون با تغییر ارزش هر صفحه، با استفاده از انتشار، ارزش کلیه صفحات دستخوش تغییر قرار می‌گیرد این الگوریتم به صورت تکراری آنقدر اجرا می‌شود تا در پایان ارزش همه صفحات به عددی همگرا شود.

رتبه صفحه (PAGE RANK)

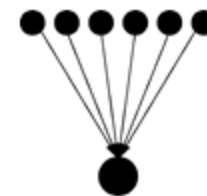
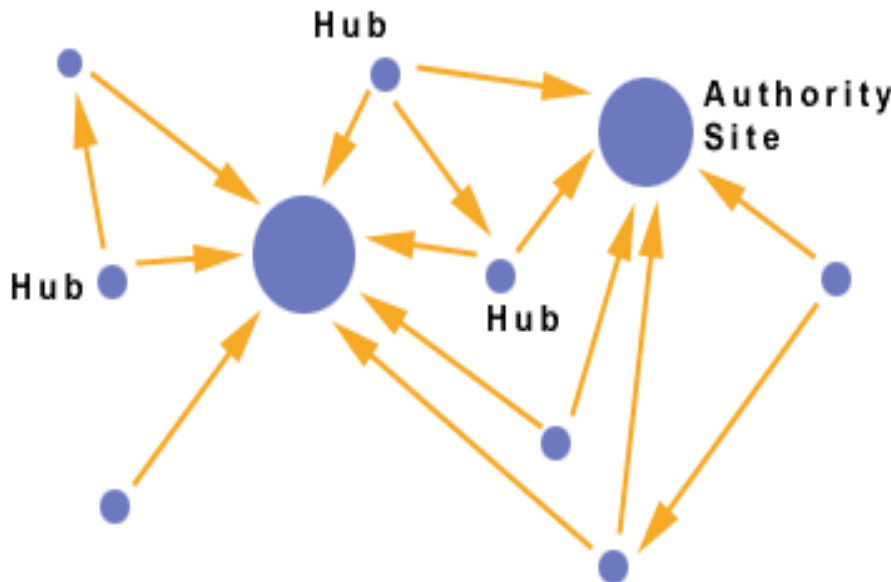
تنها گره هایی ارزش می گیرند که دارای همسایه باشند و لینکی از دیگران به آنها وارد شده باشد و در غیر اینصورت ارزش صفر را خواهند گرفت.



- گرهی دارای ارزش Hubs بالاست که به تعداد زیادی گره معتبر لینک داده باشد.



- گرهی از Authorities بالایی برخوردار است که توسط تعداد زیادی هاب لینک داده شده باشد. این مقادیر بر حسب یکدیگر و به صورت بازگشتی محاسبه می‌شوند.



○ این الگوریتم با استفاده از ساختار شبکه و پیوندهای میان گره‌ها، به ارزش‌دهی به گره‌های گراف می‌پردازد و بر دو اصل استوار است:

■ **Authorities:** صفحاتی که اطلاعات ارزشمندی دارند:

- صفحات اول سایت‌های خبری
- صفحات اول شرکت‌ها
- لینک‌های ورودی از طرف گره‌هایی که به گره‌های دیگری با تعداد ورودی بالا لینک داده باشند، قابل اعتمادتر هستند و بنابراین در مراحل بعد برای مشخص نمودن اعتبار یک گره، دیگر به همه لینک‌های ورودی به یک چشم نگاه نمی‌شود و به صورت وزندار این تعداد با هم جمع می‌شوند.

■ **Hubs:** صفحاتی که به **Authorities** لینک داده‌اند:

- شرکت‌های همکار
- هرچه گره‌ای لینک‌های ورودی بیشتری داشته باشد آن گره معتبرتر است.

CLUSTERING

خوشه بندی

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

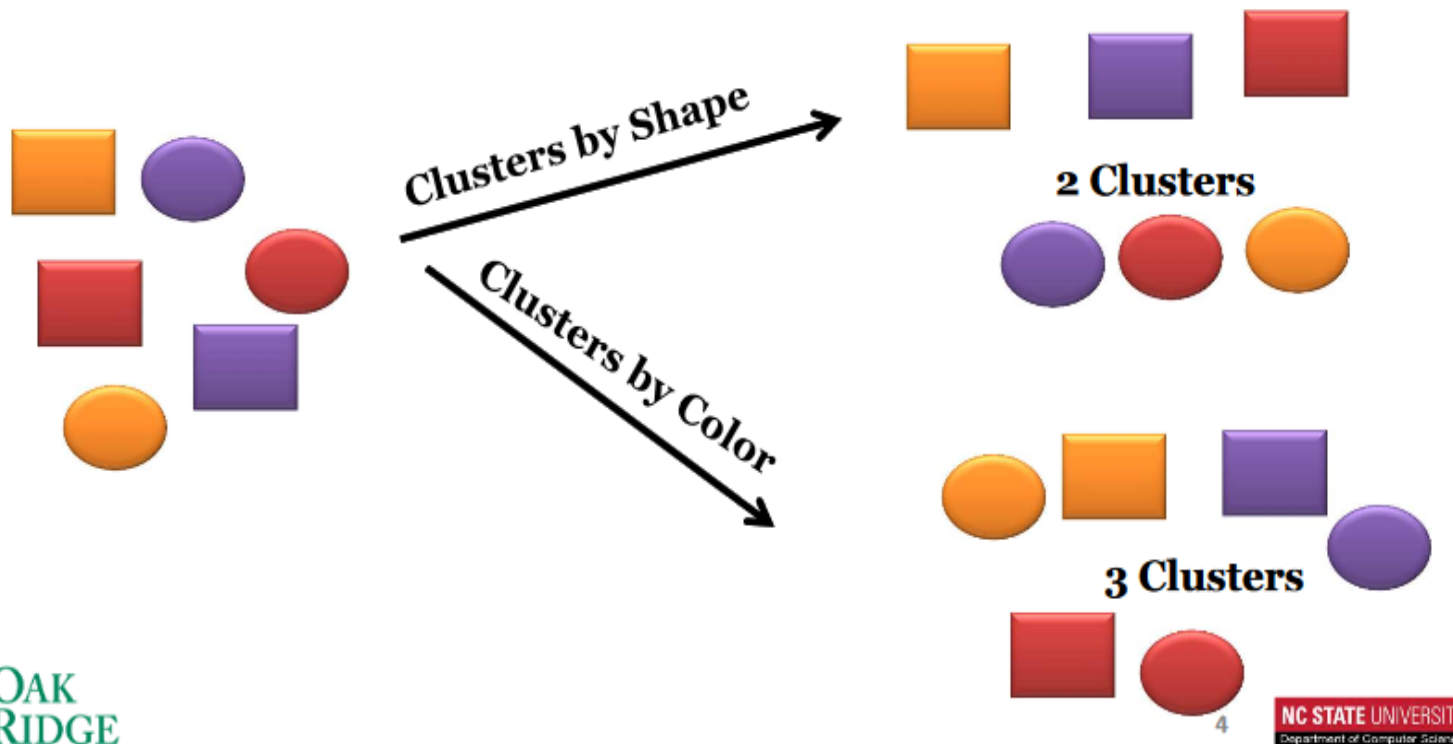
علیرضا حبیبی

خوشه بندی CLUSTERING

- هدف از خوشه بندی، استخراج بخش هایی از داده ها است که شباهت زیادی با هم دارند.
- در خوشه بندی، داده ها را به نحوی در گروه هایی تقسیم می کنیم :
 - داده های مربوط به هر گروه ویژگی های نزدیک به هم داشته باشند.
 - داده های موجود در دو گروه مختلف، ویژگی هایی با اختلاف زیاد داشته باشند.

What is Cluster Analysis?

The process of dividing a set of input data into possibly overlapping, subsets, where elements in each subset are considered related by some similarity measure



خوشه بندی و دسته بندی

○ دسته بندی:

■ ابتدا ماشین دسته‌بندی کننده با دیدن یک سری مثال و پاسخ صحیح آن‌ها (پاسخ‌هایی که توسط مربی عرضه می‌شوند) آموزش داده می‌شود و در نتیجه گروه‌هایی شکل می‌گیرند که به هر یک از آنها یک کلاس یا دسته گفته می‌شود. سپس هر یک از نمونه‌های جدید از طریق مقایسه در یکی از دسته‌های از قبل به وجود آمده قرار می‌گیرند.

○ خوشه بندی:

■ بر خلاف روش‌های بامربی، هیچ اطلاع قبلی راجع به داده‌های ورودی وجود ندارد. بنابر این، در خوشه‌بندی، گروه‌ها و تعداد واقعی آنها از قبل معلوم نیستند.

■ البته در بعضی از روش‌ها، تعداد خوشه باید در ابتدای کار توسط کاربر مشخص گردد.

○ اجزای همبند : Connected components

- اجزای همبند، زیر گراف هایی هستند که از هر نود آن زیر گراف، می توان به دیگر نودهای زیر گراف رسید.

○ زیر گراف کامل : Clique

- زیر گراف های کامل ارتباطات قوی تر میان سایت ها را مشخص خواهد کرد. زیر گراف کامل، زیر گرافی است که نودهای آن زیر گراف همگی به هم لینک مستقیم دارند (بر خلاف اجزای همبند که ممکن است نودها به هم وصل باشند اما لینک مستقیم نداشته باشند و با واسطه به هم وصل باشند).

○ K-core :

- زیر گرافی است که درجه هر گره آن حداقل k است.

○ روشهای مبتنی بر فاصله

K-MEANS ■

OPTICS ■

Hierarchical ■

○ روشهای مبتنی بر پیمانی (ماژولاریتی)

Louvain ■

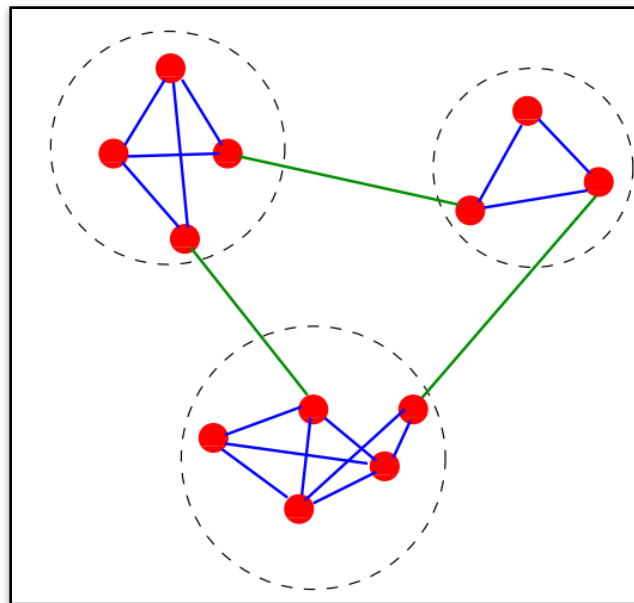
Tabrizi ■

○ انجمن های همپوشان

Label Propagation ■

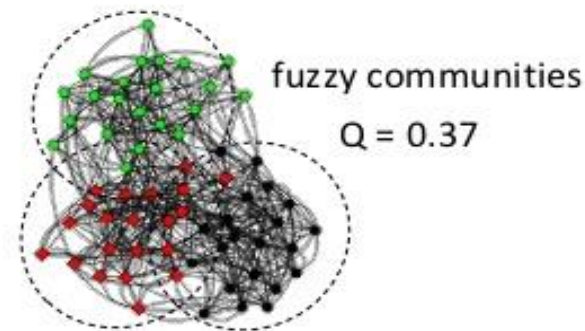
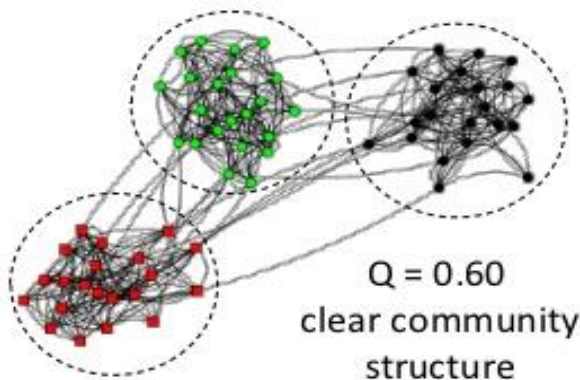
خوشه بندی

- کلیت روش‌های خوشه بندی، پیدا کردن زیرگراف‌هایی است که ارتباط درونی زیاد و ارتباطات بیرونی کمی دارند.
- با بیشینه کردن معیار پیمانی خوشه‌های گراف را پیدا کنند.



modularity - example

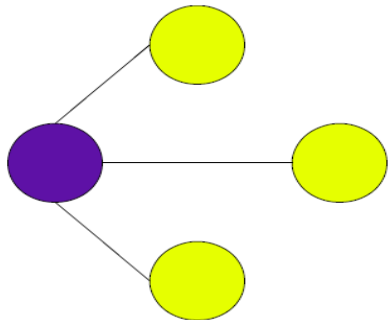
- In a random graph (ER model), we expect that any possible partition would lead to $Q = 0$.
- Typically, in non-random graphs modularity takes values between 0.3 and 0.7.



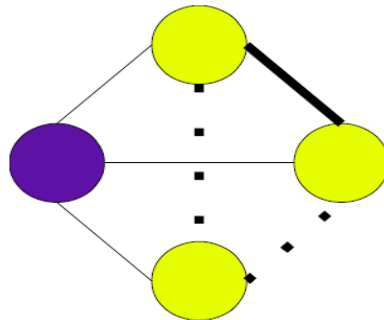
27

معیار پیمانی

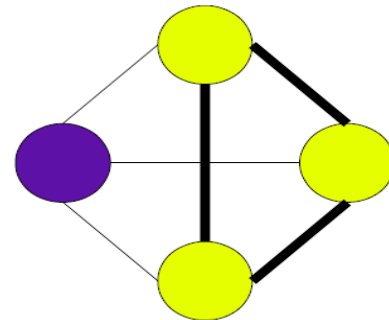
- ◉ فرم ساده شده‌ی پیمانی عبارت است از کسری از یال‌های درون یک خوشه واقعی منهای کسری از یال‌ها که در یک گراف تصادفی درون خوشه وجود می‌آید.
- ◉ اگر تعداد یال‌های درون خوشه‌ای بهتر از حالت گراف تصادفی نباشد، پیمانی برابر صفر است. حداکثر مقدار پیمانی زمانی به دست می‌آید که تمام رئوس هر خوشه به هم متصل باشند و یالی خوشه‌ها را به هم متصل نکند.
- ◉ تجربه نشان می‌دهد که در گراف‌های متعلق به شبکه‌های اجتماعی این مقدار میان ۰.۳ تا ۰.۷ است.



(a) No pairs formed among neighbors: $C = 0$



(b) One pair formed among neighbors: $C = 1 / 3$



(c) Three pairs formed among neighbors: $C = 3 / 3$

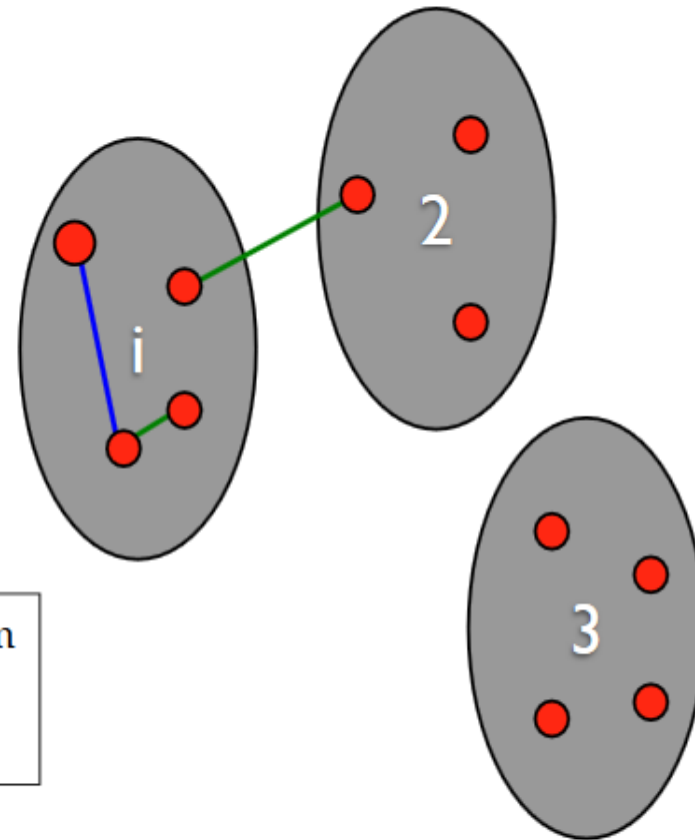
Modularity

e_{ii} = % edges in module i

$$e_{ii} = |\{(u,v) : u \in V_i, v \in V_i, (u,v) \in E\}| / |E|$$

a_i = % edges with at least 1 end in module i

$$a_i = |\{(u,v) : u \in V_i, (u,v) \in E\}| / |E|$$



Modularity is:

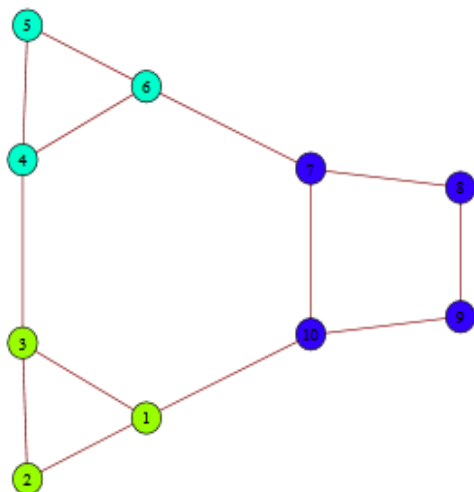
$$Q = \sum_{i=1}^k (e_{ii} - a_i^2)$$

probability a random edge would fall into module i

probability edge is in module i

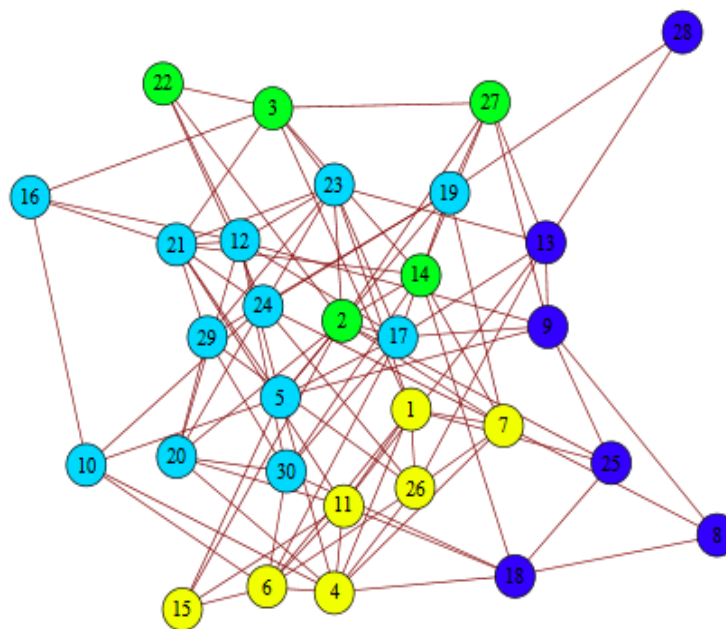
High modularity $\Rightarrow$ more edges within the module that you expect by chance.

Examples



Communities Assigned
to a small graph

Note: maximizing
modularity will find it's
own # of clusters



Communities assigned to
a random graph

LOUVAIN

◉ در این روش محدودیت سائز گراف ورودی به محدودیت حافظه ما بستگی دارد.

◉ این الگوریتم به راحتی بر گراف‌هایی با صدها میلیون راس قابل اعمال است.

◉ این الگوریتم شامل دو فاز است که به طور متناوب اجرا می‌شوند. فرض کنید که با یک شبکه وزن‌دار با n راس سروکار داریم. در ابتدا هر راس را برابر با یک انجمن می‌گیریم. بنابراین در ابتدا به تعداد رئوس، انجمن داریم.

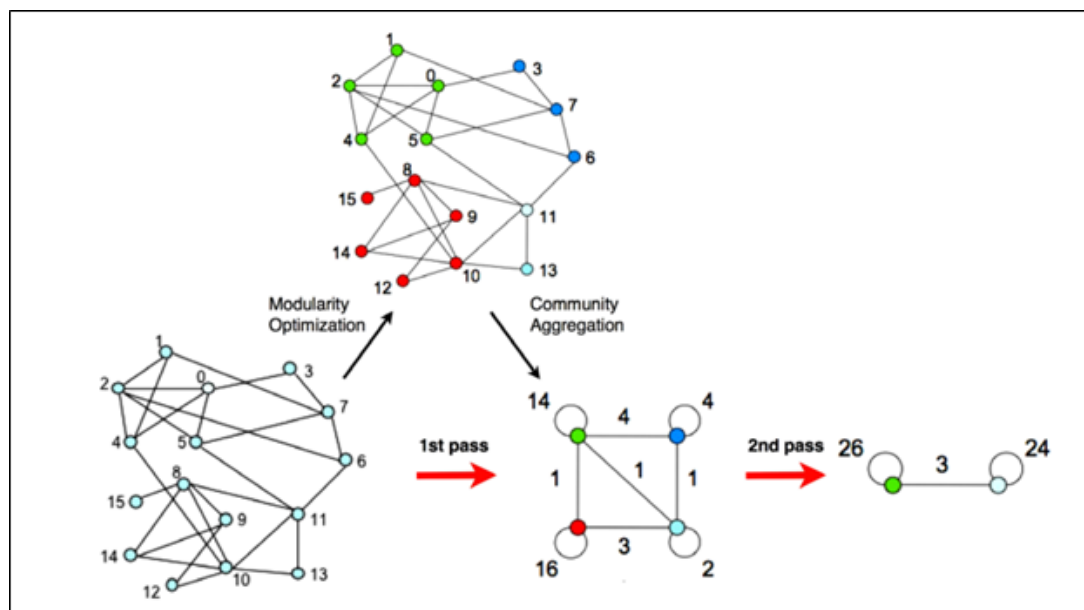
◉ فاز ۱

■ سپس برای هر راس i ، انجمن همسایه Z را به نحوی می‌یابیم که به ازای حذف i از انجمن خودش و ملحق کردن آن به انجمن Z ، پیمانی بیشینه شود. و سپس راس i را به انجمن Z اضافه می‌کنیم. این عمل تنها در صورتی انجام می‌شود که میزان پیمانی افزوده شود، در غیر این صورت راس i در انجمن خودش باقی می‌ماند.

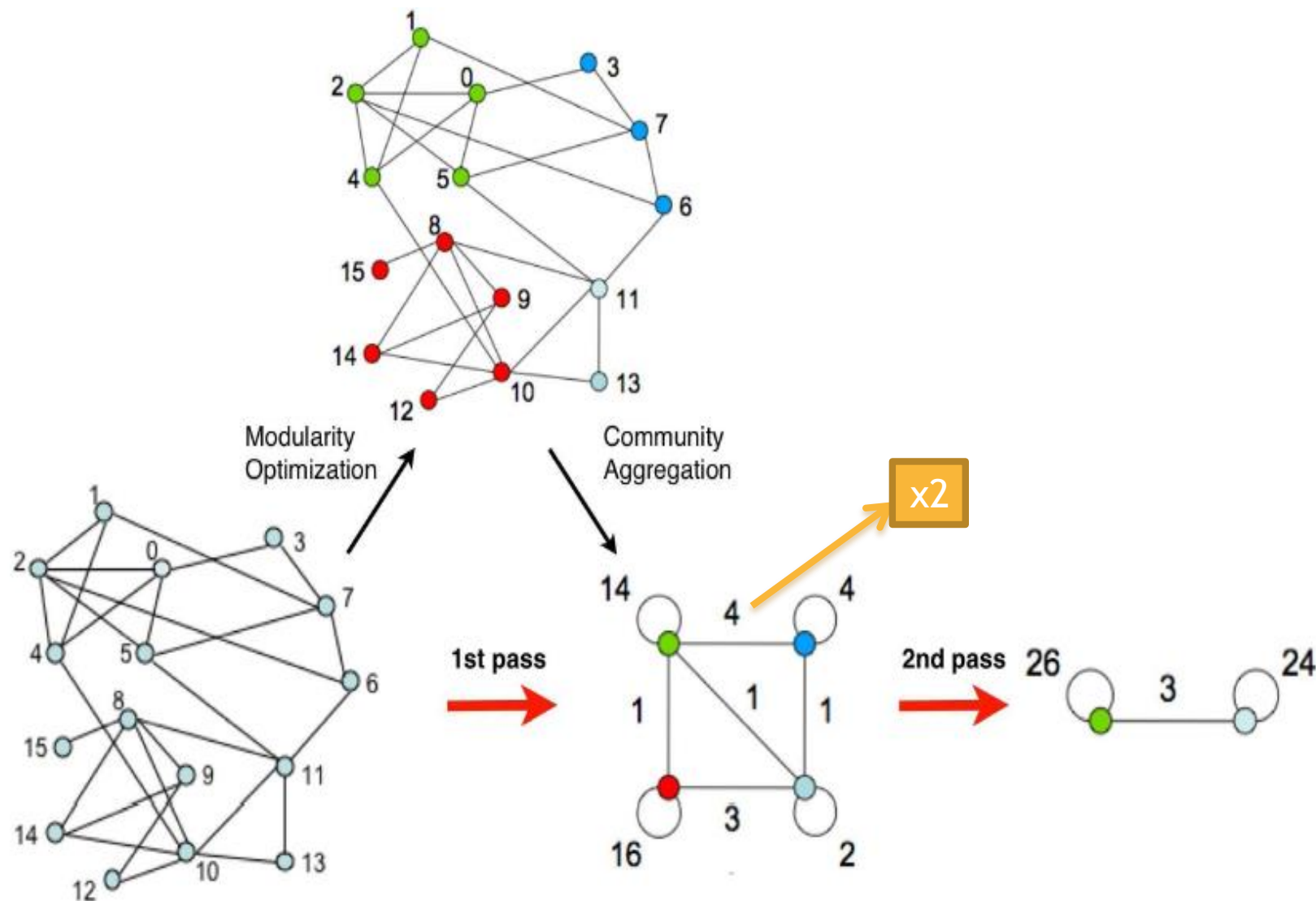
LOUVAIN

○ فاز ۲

- این فاز شبکه جدیدی را می‌سازد که رئوس آن همان انجمن‌های تشکیل شده در فاز قبل است. برای انجام این کار، وزن یال‌های میان رئوس جدید، برابر با جمع یال‌های میان رئوس درونی آنها می‌شود. ضمناً یال‌های میان رئوس موجود در یک انجمن موجب یال‌های حلقه شکل می‌گردند.
- پس از پایان فاز ۲، می‌توان مجدداً فاز ۱ را بر شبکه جدید اعمال نمود.

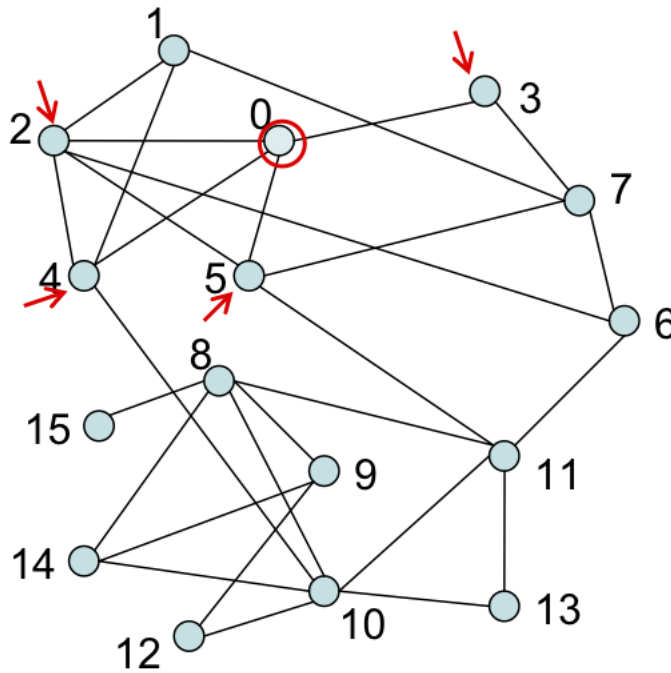


تشیص انجمن به روش LOUVAIN



An example

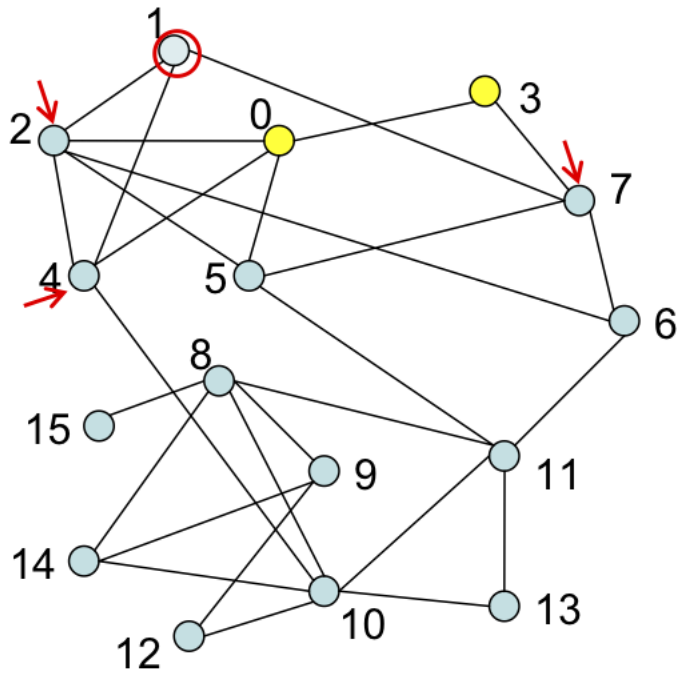
Pass 1 – Iteration 1
Each node belongs to
an atomic community



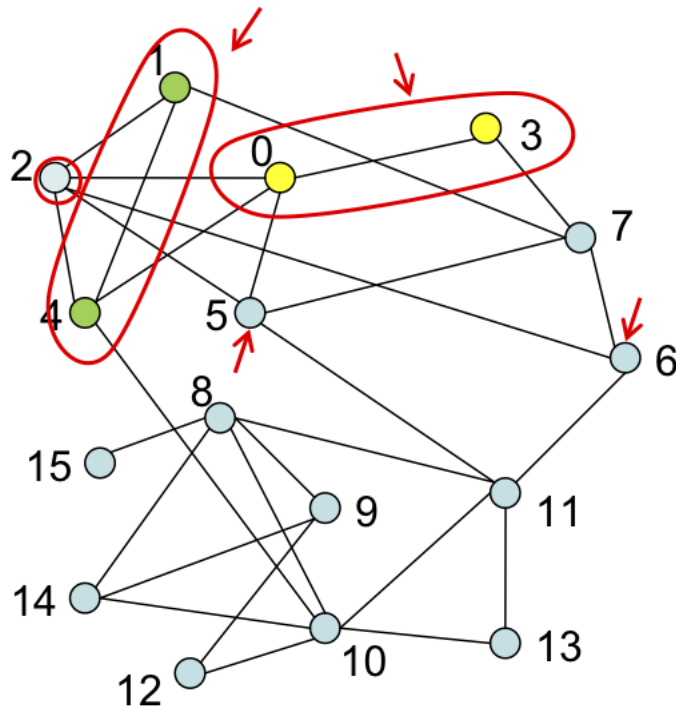
Ex: J-L Guillaume

An example

Pass 1 – Iteration 1
insert 0 in c[3]

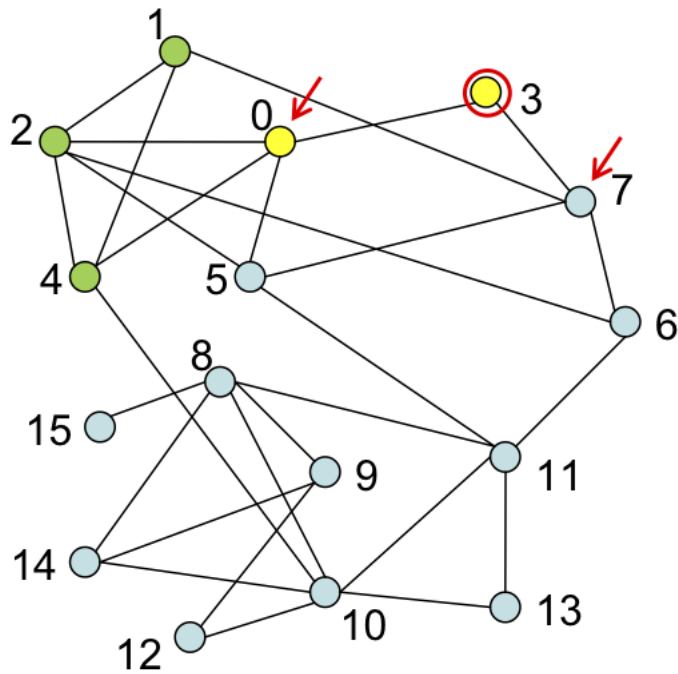


An example



Pass 1 – Iteration 1
insert 0 in $c[3]$
insert 1 in $c[4]$

An example



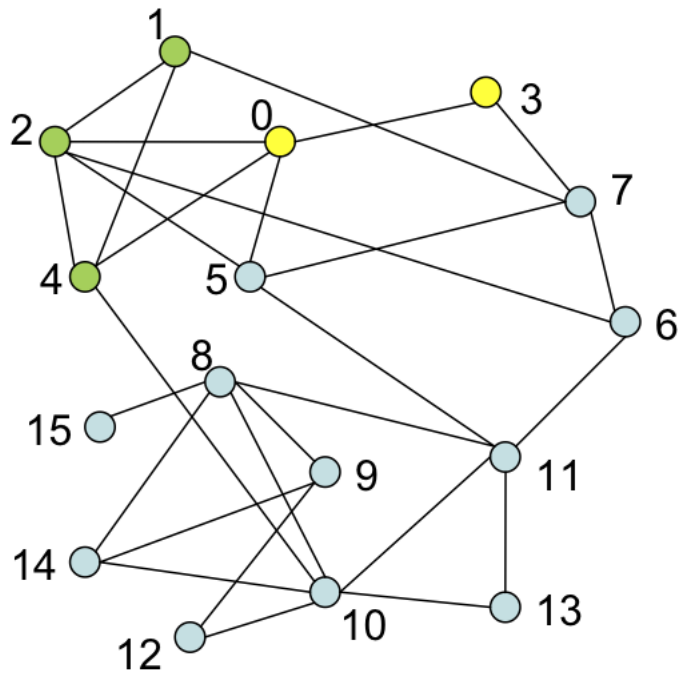
Pass 1 – Iteration 1

insert 0 in $c[3]$

insert 1 in $c[4]$

insert 2 in $c[1,4]$

An example



Pass 1 – Iteration 1

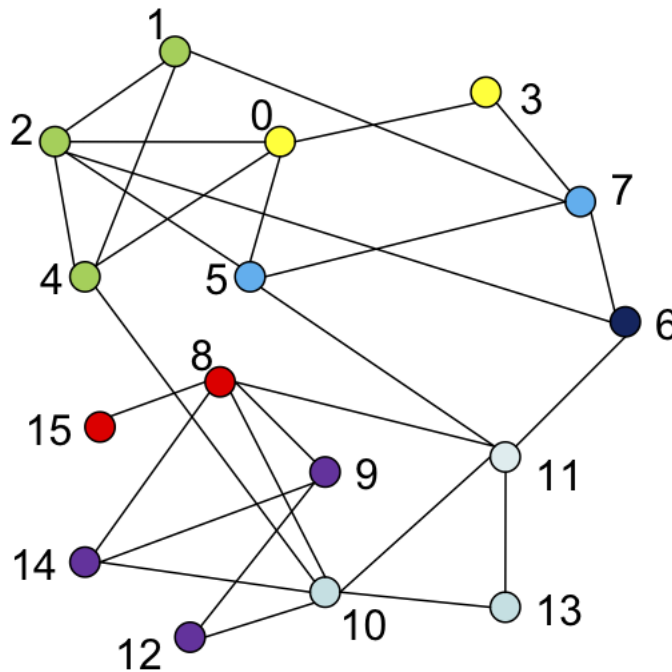
insert 0 in $c[3]$

insert 1 in $c[4]$

insert 2 in $c[1,4]$

insert 3 in $c[0]$

An example



Pass 1 – Iteration 1

insert 0 in $c[3]$

insert 1 in $c[4]$

insert 2 in $c[1,4]$

insert 3 in $c[0]$

insert 4 in $c[1]$

insert 5 in $c[7]$

insert 6 in $c[11]$

insert 7 in $c[5]$

insert 8 in $c[15]$

insert 9 in $c[12]$

insert 10 in $c[13]$

insert 11 in $c[10,13]$

insert 12 in $c[9]$

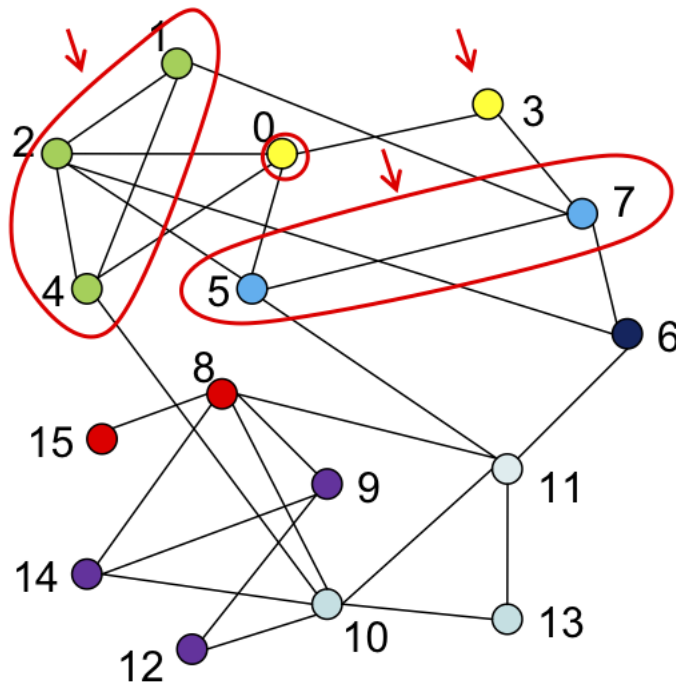
insert 13 in $c[10,11]$

insert 14 in $c[9,12]$

insert 15 in $c[8]$

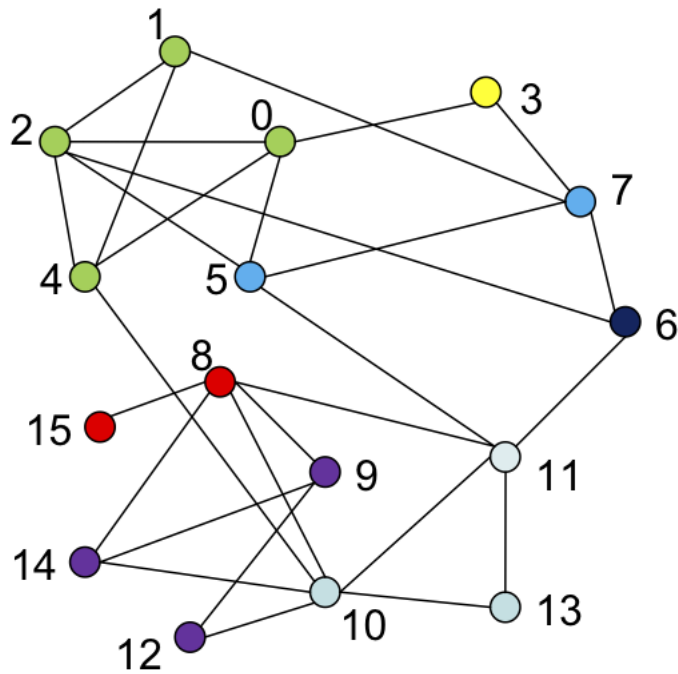
An example

Pass 1 – Iteration 2

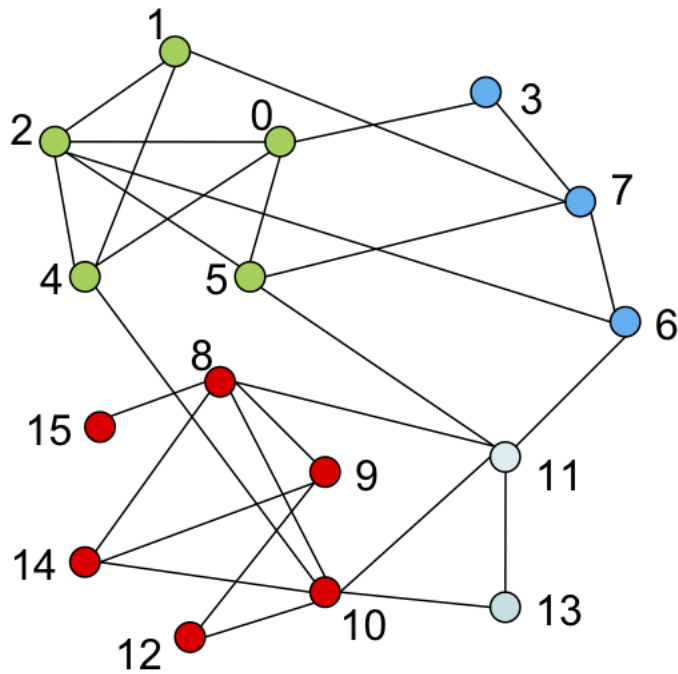


An example

Pass 1 – Iteration 2
insert 0 in c[4]

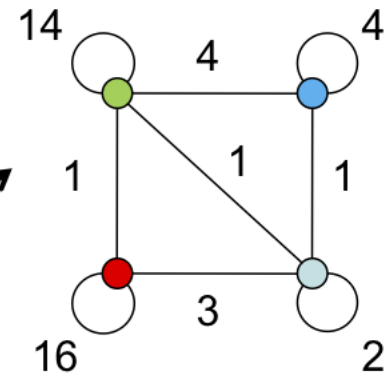


An example

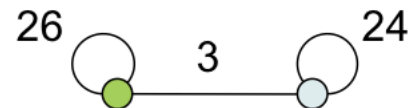


end of
pass 1

After 4 iterations



end of
pass 2



end of
pass 3

LOUVAIN

- الگوریتم بسیار سریع است. برای گراف‌های خلوت، الگوریتم مرتبه خطی دارد. دلیل سرعت بالای این روش، ساده بودن محاسبه تفاوت در میزان پیمانی و همچنین کاهش سایز گراف بعد از تعداد کمی تکرار است. به نحوی که معمولا بیشتر زمان مصرف شده، در تکرار اول الگوریتم صرف می‌شود.
- روش تبریزی : نتایجی با کیفیت لوین دارد اما بعضا در زمانی بیشترخوشه ها را تولید می کند.

روشهای مبتنی بر فاصله

- جهت استفاده از این الگوریتم‌ها برای یافتن انجمن‌ها در گراف، ماتریسی به نام ماتریس فاصله را برای گراف تعریف می‌کنیم.
- ماتریس فاصله یک ماتریس مربعی است که تعداد سطر و ستون‌هایش برابر با تعداد رئوس گراف است و هر عنصر آن، فاصله (طول کوتاه‌ترین مسیر) دو راس را نمایش می‌دهد.
- در اینجا، بردار ویژگی هر راس را برابر با سطر متناظر با آن می‌گیریم.

A	B	C	D	F	G
B	0	3			
C		0	4		
D			0		
E		6		0	
G		7			0

ماتریس فاصله: هر عنصر، طول کوتاه‌ترین مسیر بین دو گره است

روشهای مبتنی بر فاصله

○ K-Means

- این الگوریتم تمامی داده‌ها را در k خوشه گروه‌بندی می‌کند، به نحوی که هر داده به خوشه‌ای تعلق می‌گیرد که مرکزش نزدیکتر به آن داده باشد.



○ مراحل:

- (۱) k (در اینجا سه) نقطه شانسی به عنوان مرکز خوشه‌ها انتخاب می‌شوند.
- (۲) خوشه‌ها بر اساس نزدیکترین گره به مرکز خوشه‌ها شکل می‌گیرند.
- (۳) مرکز جدید هر خوشه بر اساس مرکزیت گره متعلق بر آن خوشه تعیین می‌گردند.
- (۴) مراحل ۲ و ۳ تا همگرا شدن خوشه‌ها ادامه پیدا می‌کنند.

K-MEANS

- این الگوریتم برای پردازش مجموعه داده‌های بزرگ کارایی بالایی دارد. پیچیدگی زمانی این الگوریتم برابر $O(tkn)$ می‌باشد که N تعداد داده‌ها، k تعداد خوشه‌ها و t تعداد تکرار می‌باشد.
- این الگوریتم اکثر اوقات در یک نقطه بهینه‌ی محلی متوقف می‌شود.
- تعداد خوشه‌ها در این الگوریتم باید از قبل مشخص باشد (همانند اکثر الگوریتم‌های مبتنی بر فاصله)

تنظیمات الگوریتم K-MEANS

▪ تنظیمات الگوریتم k-means

این الگوریتم تنها دارای یک پارامتر k است که با انتخاب گزینه **Setting** می‌توان این پارامتر را تنظیم کرد. مقدار پیش‌فرض این پارامتر ۵ می‌باشد. این پارامتر برابر با تعداد انجمن‌های مورد نظر است. در واقع با تنظیم این مقدار با عدد ۵، این الگوریتم کل گراف را به ۵ انجمن تقسیم خواهد کرد. پس از اجرای این الگوریتم، یک عدد با نام **Clustering(KMeans)** به **Data Laboratory** اضافه خواهد شد که در آن به ازای هر گره شناسه انجمنی که گره متعلق به آن است مشخص شده است. همچنین خروجی این الگوریتم را می‌توان در قسمت **Partition** گفی با انتخاب پارامتر **Clustering(KMeans)** بر روی گراف مشاهده کرد.

$$k = \sqrt{\frac{n}{2}}$$

OPTICS

- این روش، یک روش مبتنی بر چگالی است.
- در روش‌های مبتنی بر چگالی، خوشه‌ها به عنوان نواحی از فضای مسئله که دارای چگالی بالایی هستند در نظر گرفته می‌شوند و به وسیله نواحی با چگالی پایین از هم تفکیک می‌شوند.
- به هر شیء یک عدد چگالی نسبت داده می‌شود. این عدد چگالی با توجه به تعداد اشیای همسایه در یک شعاع خاص به دست می‌آید.
- ایده‌ی اصلی این دسته از روش‌ها به این صورت است که یک خوشه را تا زمانی که چگالی همسایگی آن از یک آستانه بیشتر نشده است، رشد دهیم.

این الگوریتم نیز دارای دو پارامتر می باشد که با انتخاب گزینه Setting می توان مقدار آنها را تنظیم کرد. پارامتر اول این الگوریتم Eps می باشد که نشان دهنده حداکثر شعاع همسایگی برای یک گره مرکزی می باشد.

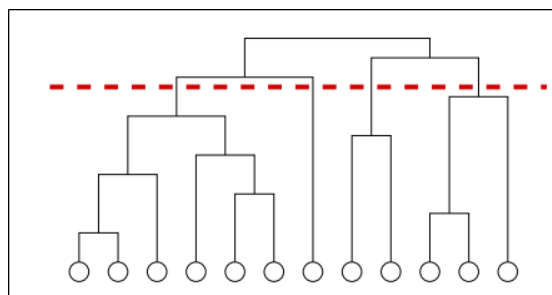
◉ پارامتر دیگر این روش MinPoints می باشد که نشان دهنده تعداد کمینه گره های موجود در شعاع همسایگی Eps برای یک گره مرکزی می باشد.

◉ با افزایش Eps در واقع به یک گره این اجازه را می دهیم که برای مرکزی بودن شعاع همسایگی بزرگتری داشته باشد. بدین ترتیب با افزایش Eps ممکن است که انجمن هایی با طول بزرگ (اعضای بیشتر) شکل گیرد. همچنین با کاهش مقدار Eps انجمن ها کوچکتر می شوند.

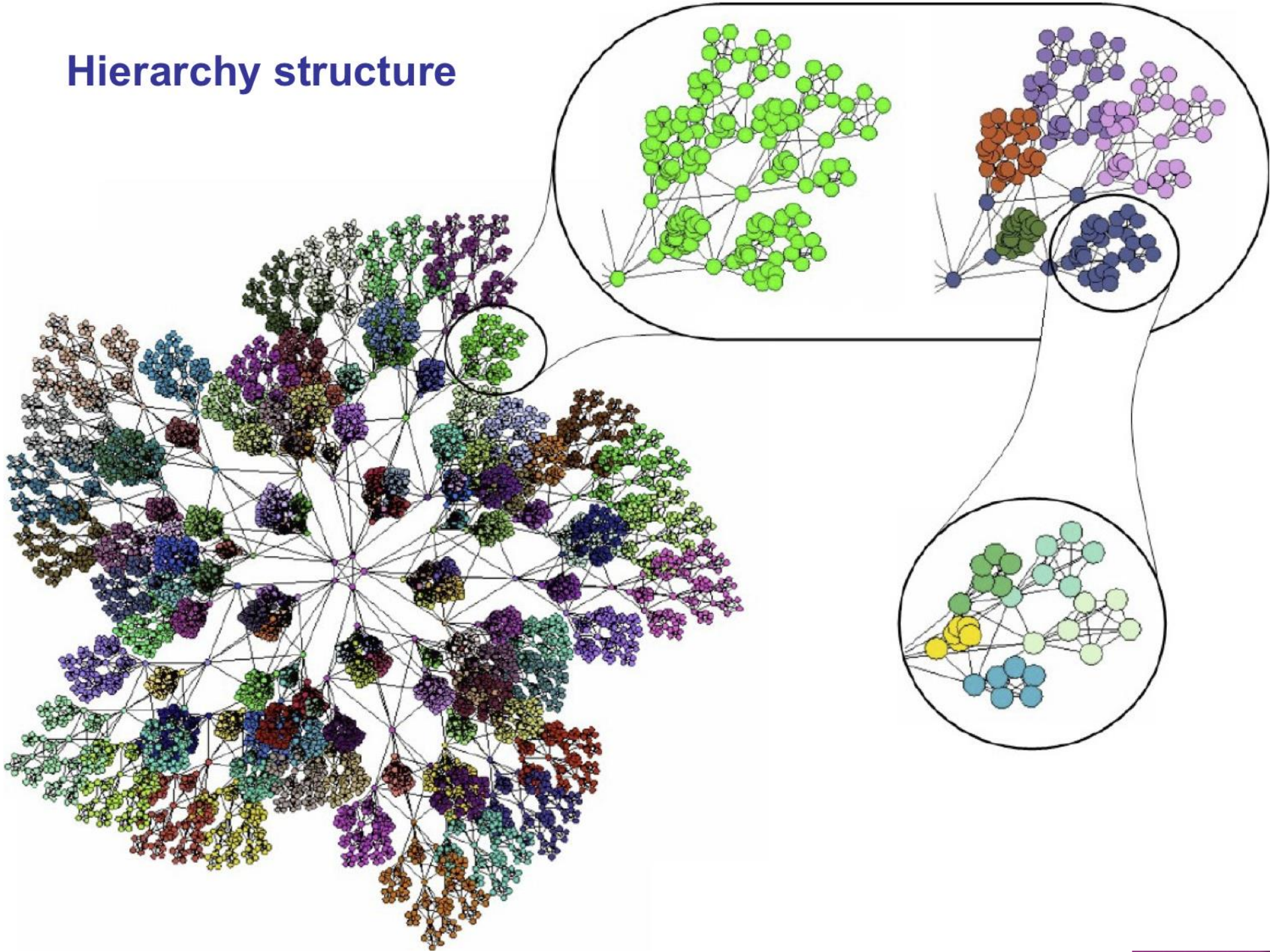
◉ همچنین با افزایش MinPoints شرط مرکزی بودن یک گره سخت تر شده و بدین ترتیب انجمن هایی با طول کوچک ممکن است شکل گیرند.

روش‌های سلسله مراتبی

- هر داده را در یک خوشه مجزا قرار می‌دهیم.
- در هر گام دو خوشه که بیشتر از همه به هم نزدیک هستند را با هم ترکیب و به یک خوشه بزرگتر تبدیل می‌کنیم.
- تا هنگامی که تمام خوشه‌ها با هم ترکیب شوند و به یک خوشه واحد تبدیل شوند، مرحله ۲ را تکرار می‌نماییم.
- پیچیدگی خوشه‌بندی تجمیعی از مرتبه‌ی $O(n^3)$ است که باعث می‌شود تا برای مجموعه‌های داده‌ای بزرگ بشدت کند باشد.



Hierarchy structure

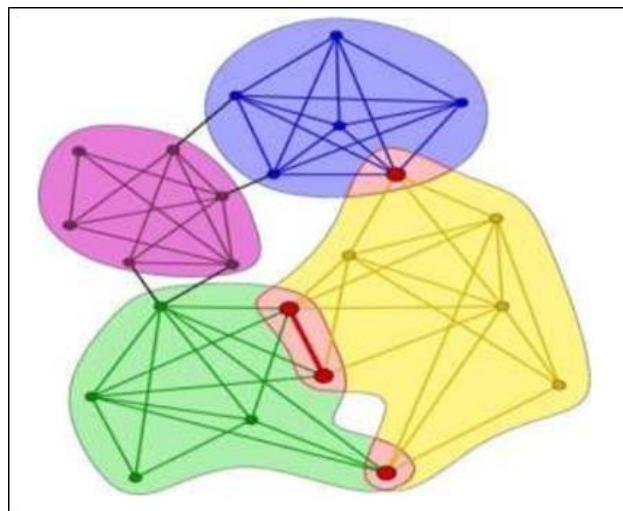


الگوریتم HIERARCHICAL

- ◉ الگوریتم تشخیص انجمن سلسله مراتبی (Hierarchical) می باشد.
- ◉ پارامتر اول این روش Number of clusters می باشد که در این تعداد انجمن هایی که الگوریتم باید تشخیص دهد را می توان مشخص کرد.
- ◉ پارامتر دیگر این روش، پارامتر Method می باشد که در آن می توان روش محاسبه فاصله میان انجمن ها را که این الگوریتم از آن استفاده می کند، مشخص کرد.
- ◉ پس از اجرای این الگوریتم، یک نام Clustering(Hierarchical) به Data Laboratory اضافه خواهد شد که در آن به ازای هر گره شناسه انجمنی که گره متعلق به آن است مشخص شده است.

یافتن انجمن‌های همپوشان

- روش‌هایی که تا اکنون به آن‌ها پرداخته‌ایم، تنها قابلیت یافتن انجمن‌های استاندارد را دارند. انجمن‌هایی که هر راس به یک و تنها یک انجمن تعلق دارد. اما در انجمن‌های واقعی یک راس می‌تواند در چندین انجمن مشترک باشد.
- مثلاً در یک خبر ممکن است که هم به گروه اقتصادی، هم به گروه اجتماعی متعلق باشد.



LABEL PROPAGATION

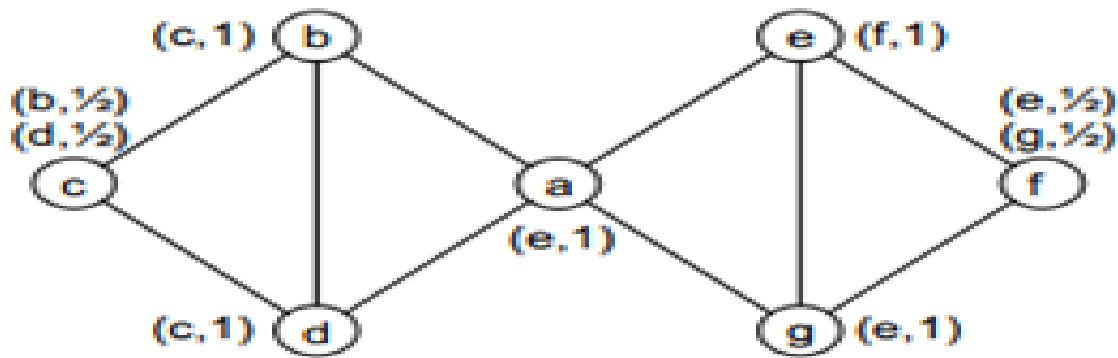
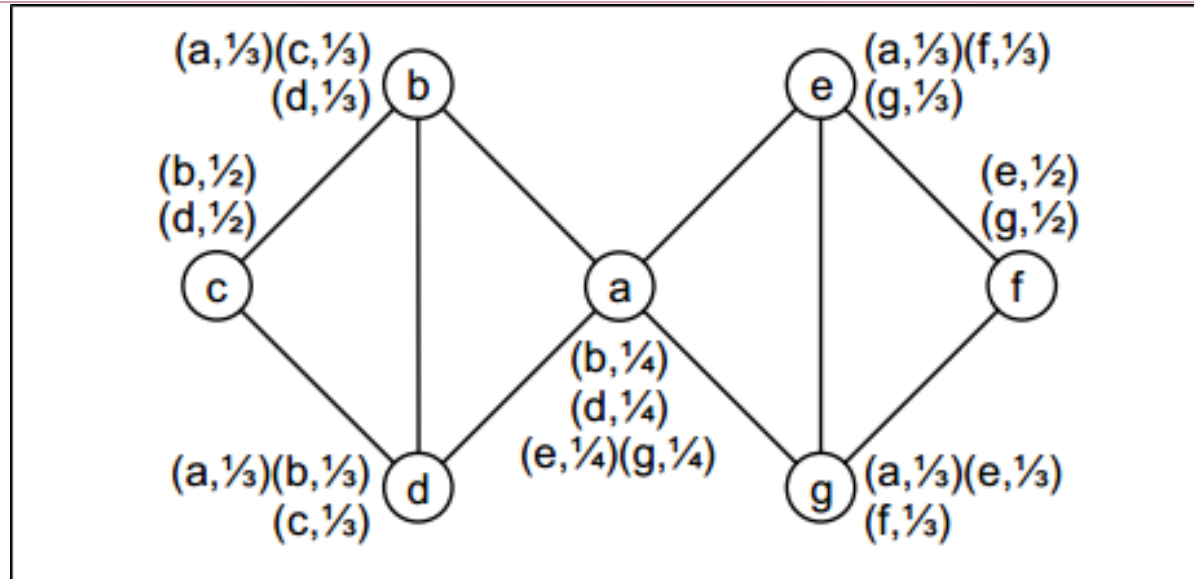
- این الگوریتم یک الگوریتم برای یافتن انجمن‌های هم‌پوشان می‌باشد. در واقع این الگوریتم این قابلیت را دارد که یک گره را در چندین انجمن مختلف قرار دهد. با انتخاب گزینه **Setting** این الگوریتم می‌توان پارامترهای آن را تنظیم کرد. پارامتر اصلی این روش V می‌باشد که نشان دهنده حداکثر تعداد انجمن‌هایی است که یک گره می‌تواند عضو آنها باشد.
- مقدار پیش‌فرض این پارامتر ۲ می‌باشد. بدین معنی که هر گره حداکثر می‌تواند بطور همزمان عضو ۲ انجمن مختلف باشد.
- پارامتر دیگر این روش **Number of threads** می‌باشد. الگوریتم **Label Propagation** بصورت چند نخ پیاده‌سازی شده است. لذا با استفاده از این پارامتر می‌توان تعداد **thread**های مورد نظر برای اجرای این الگوریتم را مشخص کرد.
- با افزایش این پارامتر سرعت اجرای الگوریتم افزایش می‌یابد. البته باید توجه داشت که با افزایش بیش از حد این پارامتر سربار ایجاد و حذف **thread**ها به برنامه اضافه می‌شود که ممکن است باعث کندی اجرای برنامه شود. پس از اجرای این الگوریتم، دو ستون به **Data Laboratory** اضافه خواهند شد.

LABEL PROPAGATION

- این روش نیز قابلیت کشف انجمن‌های همپوشان را دارد.
- این روش بر مبنای انتشار برچسب هر راس به رئوس همسایه‌اش بنا شده است و به نوعی پیمایش تصادفی گراف را در خود دارد.
- مرحله اول

- به هر یک از رئوس یک برچسب یکتا اختصاص می‌دهیم.
- در واقع برچسب هر راس نشان‌دهنده انجمن و (یا انجمن‌هایی) است که آن راس به آن تعلق دارد.
- هر راس تنها به یک انجمن تعلق دارد.
- برچسب تمامی همسایه‌های هر راس، در آن راس کپی می‌شود. همراه با هر برچسب از همسایه‌ها، عددی نیز ذخیره می‌گردد که نشان‌دهنده‌ی تعلق میزان وابستگی یک راس به یک انجمن را نشان می‌دهد.
- در هر تکرار، برای هر راس تنها انجمن‌هایی را به عنوان برچسب نگه می‌داریم که میزان ضریب تعلق آنان بیشتر از یک میزان آستانه باشد، و باقی را حذف می‌نماییم.
- ۷ نشان‌دهنده حداکثر تعداد انجمن‌هایی است که یک راس می‌تواند عضو آنان باشد.

LABEL PROPAGATION



V=2

LABEL PROPAGATION

- ◉ مرتبه زمانی این الگوریتم $O((V+E)(E/V^2))$ است.
- ◉ این الگوریتم برای گراف های که چگالی آن ها زیاد است، مناسب نمی باشد.

این روش ابتدا هر گره را به عنوان یک انجمن در نظر می‌گیرد. سپس دو انجمنی که بیشترین شباهت را با هم دارند ادغام شده و تشکیل یک انجمن جدید می‌دهند. فاصله بین دو انجمن در این روش با استفاده از رابطه زیر محاسبه می‌شود:

$$D(X, Y) = \min_{x \in X, y \in Y} d(x, y) \quad \odot$$

که در این فرمول، X ، Y نشان دهنده انجمن‌ها، x و y نشان دهنده اعضای دو انجمن هستند. و $d(x, y)$ فاصله این دو عضو را نشان می‌دهد. در واقع در این روش شباهت دو انجمن برابر با فاصله دو عضوی از دو انجمن در نظر گرفته می‌شود که در مقایسه با سایر اعضای دو انجمن کمترین فاصله را با یکدیگر دارند. در هر مرحله دو انجمنی که کمترین فاصله را با یکدیگر داشته باشند با هم ادغام می‌شوند.

این روش برعکس روش قبلی عمل می‌کند و فاصله دو انجمن را برابر فاصله دورترین جفت گره‌ها از دو انجمن در نظر می‌گیرد.

$$D(X, Y) = \max_{x \in X, y \in Y} d(x, y)$$

که در این فرمول، X, Y نشان دهنده انجمن‌ها و x, y نشان دهنده اعضای دو انجمن هستند. $d(x, y)$ نیز فاصله این دو گره را نشان می‌دهد. در این روش نیز در هر مرحله دو انجمنی که بیشترین فاصله بین دو گره از این انجمن‌ها از یکدیگر کمینه باشد، با یکدیگر ادغام می‌شوند.

این روش میانگین فاصله اعضای دو انجمن را به عنوان فاصله دو انجمن در نظر می گیرد. فاصله دو انجمن از یکدیگر با استفاده از این روش توسط رابطه زیر محاسبه می شود:

$$D(X, Y) = \frac{1}{|X| |Y|} \sum_{x \in X} \sum_{y \in Y} d(x, y) \quad \odot$$

که در این فرمول نیز X و Y نشان دهنده ی انجمن ها ، x و y نشان دهنده ی اعضای دو انجمن و $d(x, y)$ فاصله این دو گره را نشان می دهد. همچنین $|X|$ و $|Y|$ تعداد اعضای انجمن ها را نشان می دهند. در این روش نیز در هر مرحله دو انجمنی که میانگین فاصله بین گره های آنها کمینه است با یکدیگر ادغام می شوند. این روش تمایل دارد تا خوشه هایی با قطر یکسان را بیابد.

پارامتر دیگر این روش Community Limit می باشد که در با استفاده از آن می توان حداکثر تعداد انجمن هایی که باید الگوریتم آنها را پیدا کند، مشخص کرد. با عنوان مثال اگر مقدار این پارامتر ۵ باشد، الگوریتم گراف را حداکثر به ۵ انجمن تقسیم می کند.

مقدار پیش فرض این پارامتر صفر می باشد که بدین معنی است که محدودیت خاصی برای تعداد انجمن ها گذاشته نشده است. پارامتر دیگر این روش پارامتر Number of threads می باشد.

مشابه روش Label Propagation، الگوریتم Tabrizi نیز بصورت multi-thread و موازی پیاده سازی شده است و با استفاده از این پارامتر می توان تعداد thread های مورد نظر برای اجرای این الگوریتم را مشخص کرد.

پس از اجرای این الگوریتم، یک با نام Clustering(Tabrizi) به Data Laboratory اضافه خواهد شد که در آن به ازای هر گره شناسه انجمنی که گره متعلق به آن است مشخص شده است. همچنین خروجی این الگوریتم را می توان در قسمت Partition گفی با انتخاب پارامتر Clustering(Tabrizi) بر روی گراف مشاهده کرد.

▪ تنظیمات الگوریتم Tabrizi

این روش نیز دارای چندین پارامتر است که در Setting آن قابل تنظیم هستند. با استفاده از پارامتر Use Weights می‌توان مشخص کرد که الگوریتم از وزن یال‌ها در فرایند تشخیص انجمن استفاده کند یا خیر. در حالت کلی استفاده از وزن‌ها می‌تواند موجب افزایش دقت روش شود. پارامتر Number of trials تعداد قدم‌زن‌های تصادفی برای محاسبه PageRank را مشخص می‌کند. در حالت کلی تعداد بالای قدم‌زن‌های تصادفی می‌تواند دقت روش را بالا ببرد ولی تعداد زیاد این قدم‌زن‌ها می‌تواند موجب کندی اجرای الگوریتم نیز شود.

پارامتر دیگر این روش Number of trial multiplicands می‌باشد. این پارامتر نیز به منظور تنظیم تعداد قدم‌زن‌های تصادفی در نظر گرفته شده است. تعداد قدم‌زن‌های تصادفی در نهایت بصورت بیشترین مقدار از میان پارامترهای Number of trials و ضرب درجه هر گره در پارامتر Number of trial multiplicands مشخص می‌شود.

پارامتر دیگر این روش Alpha value of pagerank می‌باشد که با استفاده از آن می‌توان مقدار پارامتر آلفا (احتمال پرش تصادفی از یک گره به گره دیگر در گراف) روش PageRank را که این الگوریتم از آن استفاده می‌کند، تنظیم کرد.

نتیجه گیری

در این بخش کارایی روشهای مختلف خوشه بندی بر روی مجموعه های دادهای مختلف با ویژگیهای مختلف مورد ارزیابی قرار گرفت. شبکه های استفاده شده برای آزمایش روشهای خوشه بندی را میتوان به سه دسته تقسیم کرد:

- شبکه های چگال

- شبکه های با قابلیت خوشه بندی بالا

- شبکه های دارای یک مولفه عظیم

کارایی روشهای خوشه بندی بسته به نوع شبکه مورد استفاده میتواند متفاوت باشد.

روشهای مبتنی بر معیار پیمانی بر روی هر سه دسته از این مجموعه های دادهای کارایی بالایی دارند و خوشه های موجود در این مجموعه های دادهها را با دقت مناسبی پیدا میکنند.

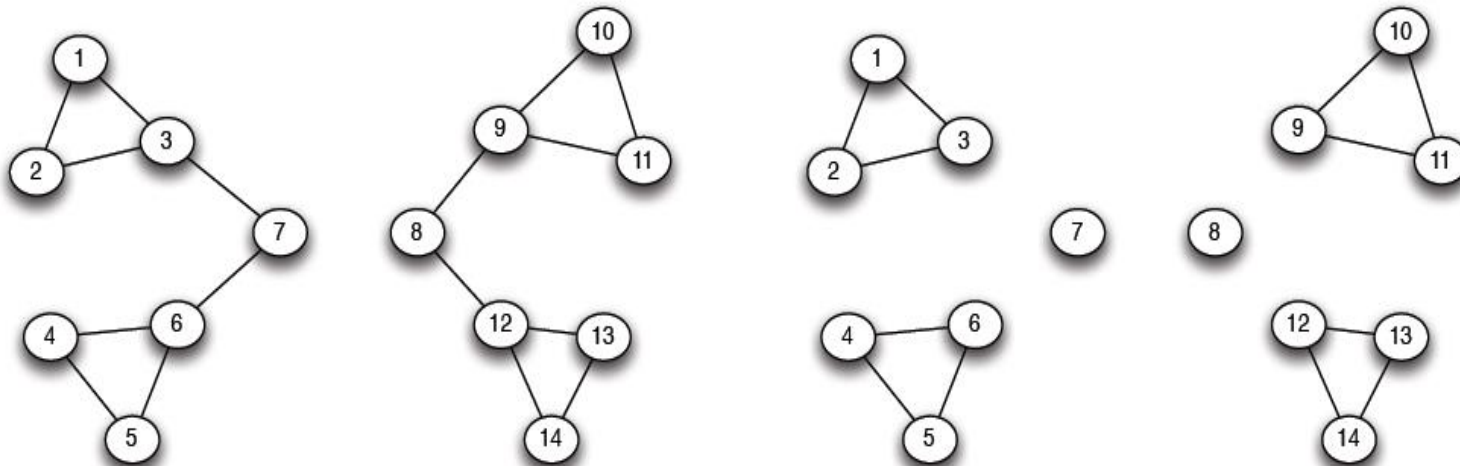
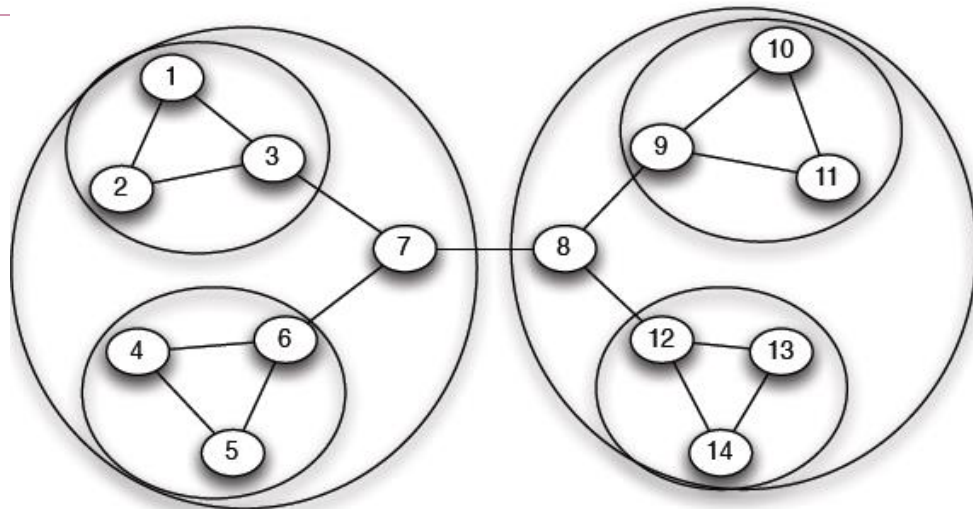
روشهای مبتنی بر Random walk (Label propagation) بر روی مجموعه های دادهای چگال کارایی چندان مناسبی ندارند.

زیرا گره های موجود در این گرافها جدایی چندانی از هم ندارند. بر خلاف گرافهای چگال، این روشها بر روی سایر گرافها کارایی بسیار بالایی دارند و به خوبی میتوانند خوشه های موجود در این گرافها را بیابند.

نتیجه گیری

- روش Label propagation در تشخیص خوشه های موجود در گراف با مولفه عظیم نیز کارایی بسیار مناسبی دارد. همچنین یکی دیگر از ویژگیهای این روش قابلیت آن در تشخیص خوشه های همپوشان میباشد.
- در واقع این روش قادر است تشخیص دهد که آیا یک گره در چندین خوشه عضو است یا خیر. این قابلیت در بسیاری از کاربردها میتواند مفید باشد.
- روشهای مبتنی بر فاصله و روش سلسه مراتبی از دقت مناسبی در تشخیص خوشه های مربوط به گرافهای با قابلیت خوشه بندی بالا برخوردارند.
- مشکل اصلی این روشها زمان اجرای بسیار بالای آنها میباشد. این روشها در خوشه بندی گرافهای با تعداد گره های زیاد کارایی بسیار پایینی از نظر زمان اجرا دارند.

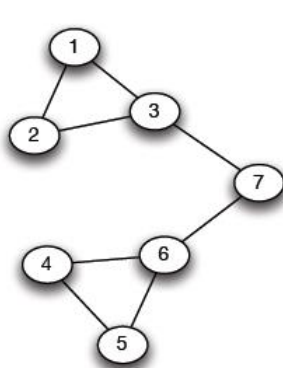
تشخیص انجمن با استفاده از بینابینی یالها (روش نیومن)



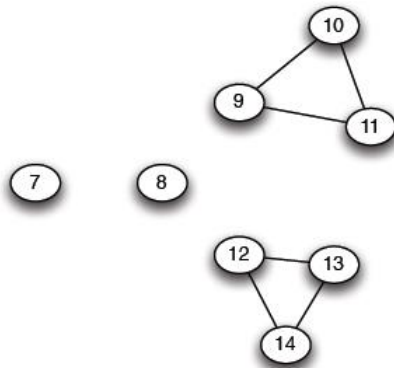
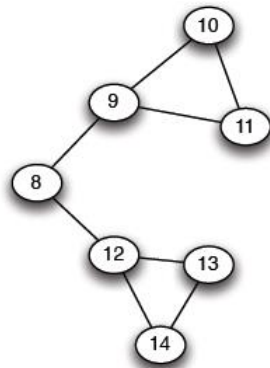
(a) Step 1

(b) Step 2

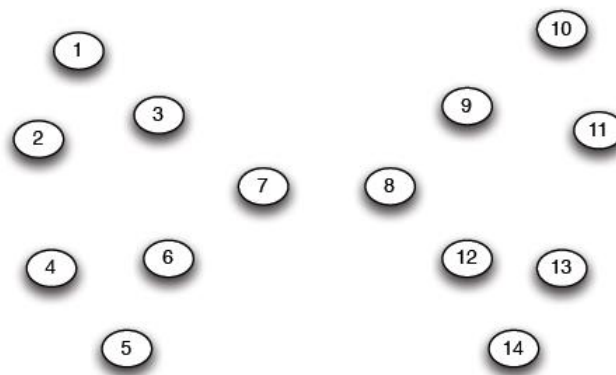
روش نیومن



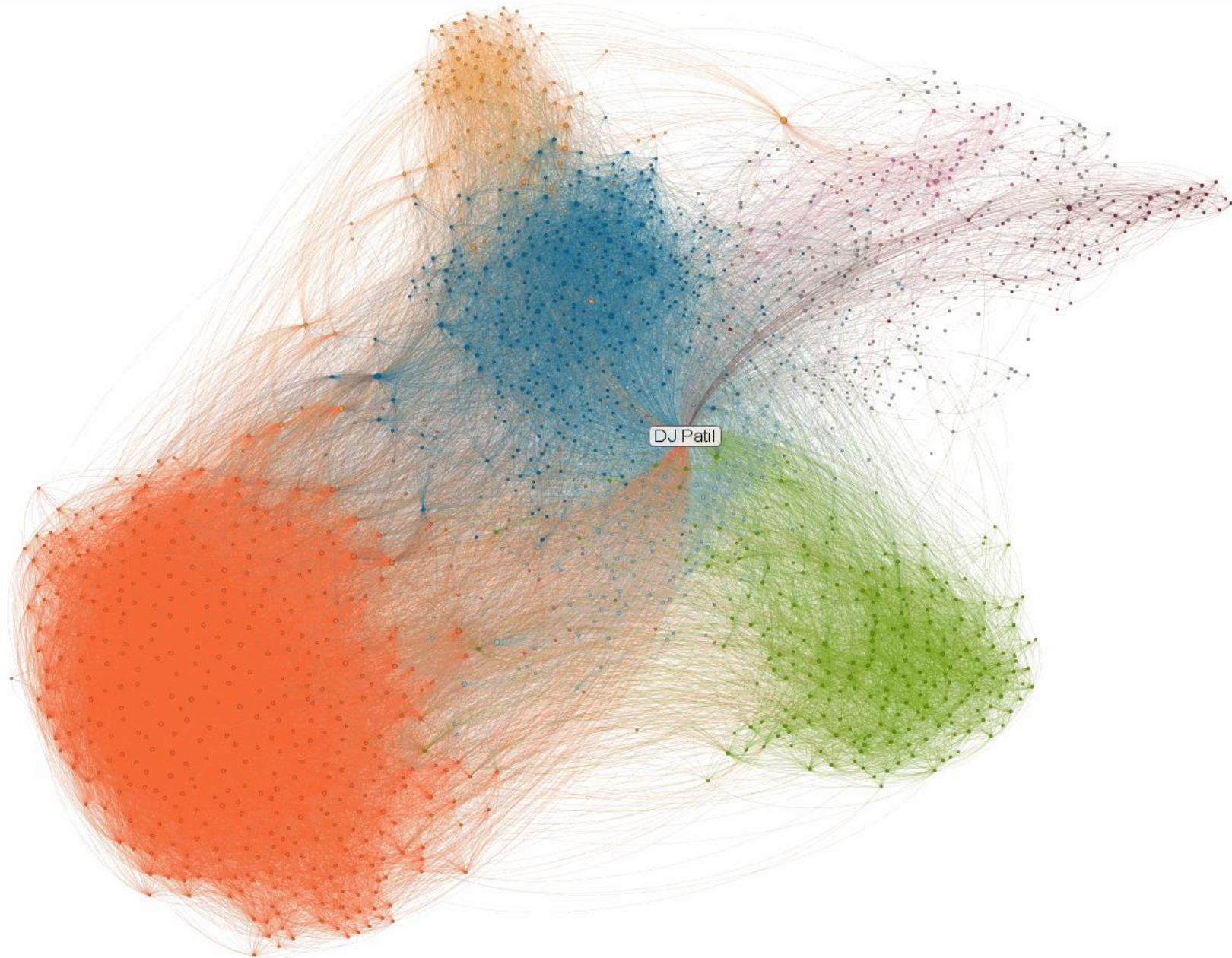
(a) Step 1



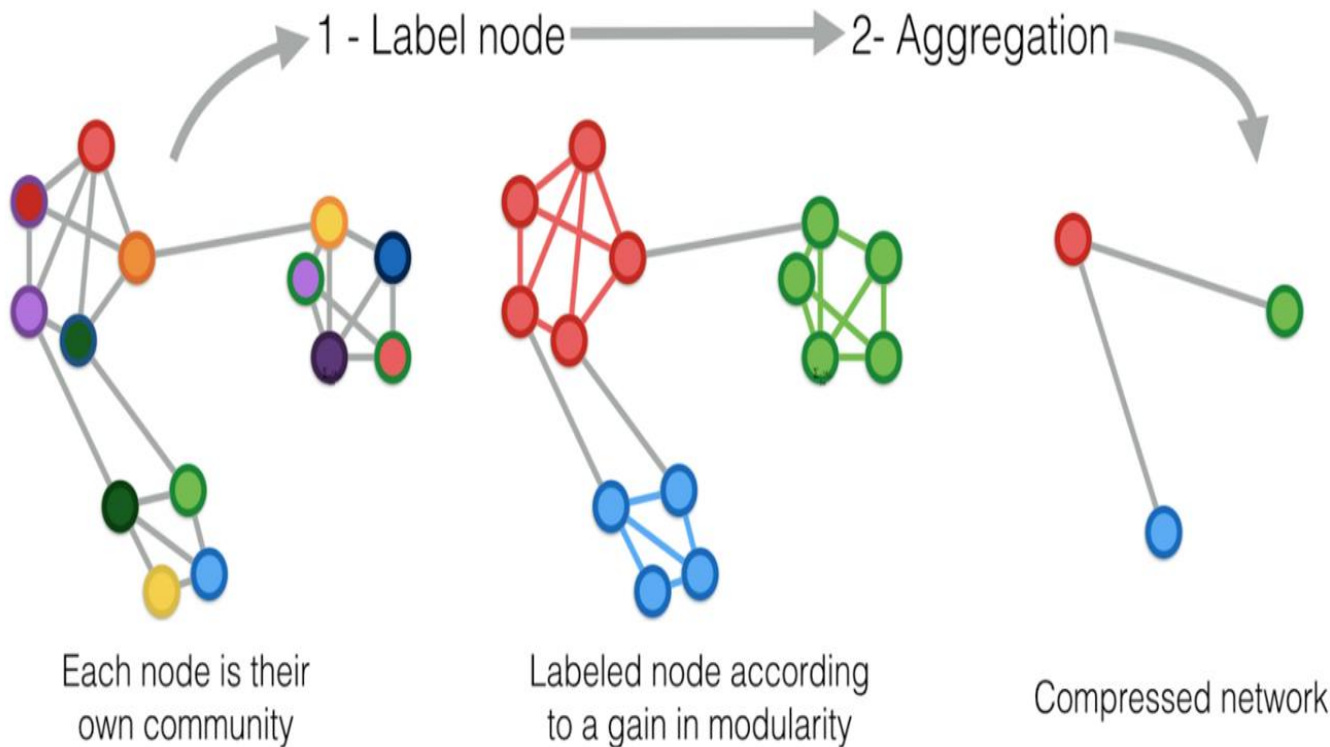
(b) Step 2



(c) Step 3



AGGREGATION



LINK PREDICTION

پیشبینی لینک

www.bigdata-ir.com

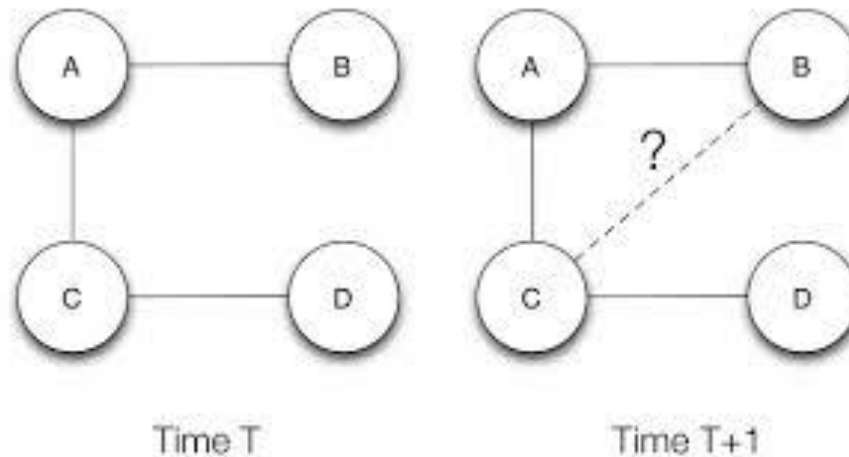
Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

پیش بینی لینک

به چه احتمالی بین دو گره یک لینک ارتباطی به وجود خواهد آمد؟



پیش بینی لینک

LINK PREDICTION

- پیش بینی وجود ارتباط میان دو موجودیت بر اساس ویژگی های موجودیت ها و دیگر لینک های مشاهده شده در گراف را پیش بینی لینک می گویند .
- در حوزه پزشکی می توان شیوع یک بیماری خاص و بررسی اثر واکنش تولید شده را بر روی آن به کمک پیش بینی لینک مورد مطالعه قرار داد.

○ روش‌های ساختاری و بدون ناظر

- تنها از یک تصویر لحظه‌ای گراف شبکه استفاده می‌کنند.
- محتوای شبکه را در نظر نمی‌گیرند.
- تنها به کمک ویژگی‌های ساختاری گراف شبکه سعی در پیش‌بینی لینک‌ها در تصویر لحظه‌ای بعدی از شبکه می‌نمایند.
- سرعت مناسبی دارند.
- دقت خروجی این دسته از روش‌ها نسبت به روش‌های پیش‌بینی لینک با ناظر کمتر می‌باشد.

● مبتنی بر همسایه مشترک

- Adamic and Adar
- Resource Allocation
- شباهت کسینوسی Cosine Similarity
- SimRank
- CommonNeighbor
- ضریب جاکارد Jaccard's Coefficient
- اندیس سورنسن Sørensen Index

● مبتنی بر مسیر یا فاصله Distance (فاصله کمتر احتمال ارتباط بیشتر)

- روش های مبتنی بر گام تصادفی
- Rooted PageRank
- Katz

● مبتنی بر درجه (درجه های بیشتر احتمال ارتباط بیشتر)

- Attachment Preferential

○ بر حسب همسایه‌های یک گره میزان شباهت آن با سایر گره‌ها محاسبه می‌شود.

○ اساس این روش برپایه ویژگی تراگذاری در شبکه‌های اجتماعی است.

○ همسایه‌های مشترک **Common Neighbors**

■ امتیاز یک لینک (جدید) برابر با تعداد همسایه‌های مشترک میان نقاط انتهایی آن خواهد بود.

○ ضریب جاکارد **Jaccard's Coefficient**

■ معیار سنجش شباهت، ضریب جاکارد می‌باشد

■ ضریب جاکارد می‌گوید که دو گره شبیه‌تر به یکدیگر هستند اگر تعداد همسایه‌های مشترک زیاد و تعداد همسایه‌های غیر مشترک کمی داشته باشند.

■ اگر ضریب جاکارد را به عنوان معیاری برای رتبه‌بندی یال‌ها استفاده کنیم، به نوعی حالت نرمال شده روش همسایه‌های مشترک خواهد بود.

○ اندیس سورنسن **Sørensen Index**

■ این اندیس بیشتر برای جوامع زیست محیطی به کار می‌رود.

آدامیک آدار

ADAMIC AND ADAR

- ◉ آدامیک و آدار در زمینه وب برای تشخیص شباهت دو وب سایت به یکدیگر معیاری را به نام آدامیک آدار را معرفی نمودند
- ◉ پس از بدست آوردن ویژگی‌های مشترک میان دو صفحه در وب، به ویژگی‌هایی که نادر بوده و تنها میان دو صفحه مورد نظر به وجود آمده است، وزن بالاتری داده می‌شود.
- ◉ برای پیش‌بینی لینک می‌توان ویژگی مشترک را همسایه‌های دو گره در نظر گرفت. حال همسایه‌های مشترکی ارزشمندتر هستند که درجه کمتری دارند. به عبارت دیگر گره‌ای خاص میان دو گره مورد نظر ارزش بیشتری نسبت به گره‌هایی که دارای درجه‌ی بالایی هستند، دارد.
- ◉ فرض کنید گراف مورد نظر، گراف نویسندگان همکار باشد. منطقی به نظر می‌رسد که اگر نویسنده‌ای دارای همکارهای زیادی باشد، احتمال آنکه مشترک میان دو گره خاص باشد نیز زیادتر خواهد بود؛ ولی این احتمال برای گره‌های با درجه کمتر ناچیز می‌باشد. فرد واسطی که میان دو نفر که تنها دوستان وی می‌باشند، نقش بیشتری در نزدیک کردن آن دو گره به یکدیگر ایفا می‌کند.
- ◉ معیار آدامیک آدار یکی از روش‌های برتر در میان روش‌های پیش‌بینی لینک ساختاری بدون ناظر می‌باشد.

اختصاص منابع

RESOURCE ALLOCATION

- بسیار شبیه به روش آدامیک آدار می باشد
- همسایه های مشترک دارای درجه ی بالا را اثر دهی کمتری نسبت به روش آدامیک آدار در تعیین امتیاز لینک ها می دهد.

شباهت کسینوسی

COSINE SIMILARITY

این روش یکی از اشکال مختلف نرمال سازی فرمول محاسبه‌ی همسایه‌های مشترک به حساب می‌آید.

SIMRANK

○ سیمرنگ:

○ معمولاً در پیش‌بینی لینک دو موجودیت را که دارای ویژگی‌های مشترکی هستند را بیشتر شبیه به یکدیگر دانسته و احتمال بیشتری برای وقوع لینک میان آن‌ها متصور می‌شود.

○ از دیدگاه پیش‌بینی لینک، دو گره شانس بیشتری برای داشتنی لینک دارند اگر هر دو به تعداد زیادی گرهی مشابه متصل باشند.

روش های مبتنی بر فاصله

DISTANCE

- زوج گره‌ای که بین آن‌ها لینکی وجود ندارد و در فاصله کمی از یکدیگر قرار دارند، بیشترین احتمال برای به وجود آمدن ارتباط میان آن دو نسبت به سایرین وجود دارد.
- همان مثلث ها یا سه تایی ها

روش های مبتنی بر مسیر

این روش ها بر اساس معیار مسیر، امتیاز لینک ها را مشخص می نمایند. این روش ها به نوعی ایده ی کوتاه ترین مسیر میان دو گره را به سمت محاسبه ی تمامی مسیرهای میان دو گره ارتقاء داده اند.

روش های مبتنی بر گام تصادفی نیز جزو این دسته از روش ها محسوب می شوند.

Katz

■ معیار کتز یک فرمول برای استفاده از تعداد و طول مسیرهای میان دو گره برای پیش بینی لینک است.

■ مسیرها نیازی نیست که کوتاه ترین مسیر باشند.

■ برای کاهش اثر مسیرهای طولانی در تعیین امتیاز یک لینک، از یک ضریب به نام آلفا در این فرمول استفاده شده است. این کار باعث می شود که پیش بینی لینک میان گره ها در نواحی پر جمعیت اطراف آنها نسبت به سایر نقاط بیشتر باشد.

■ اگر مقدار این ضریب خیلی کم باشد، معیار کتز شبیه همسایه های مشترک می شود.

روش‌های مبتنی بر گام تصادفی

- پیج‌رنک ریشه‌دار: Rooted PageRank
- اگر این الگوریتم را بر روی یک گراف اجرا کنیم، در نهایت یک سری احتمالات به هر گره‌ای داده می‌شود. این احتمالات، احتمال رسیدن یک گام‌بردار تصادفی به هر گره را نشان می‌دهد.
- برای استفاده از پیج‌رنک برای پیش‌بینی لینک، از این فرض استفاده می‌کنیم که اگر دو گره با اهمیت را در نظر بگیریم، احتمال زیادی وجود دارد که این دو با یکدیگر در آینده لینک داشته باشند.
- هر چقدر مقدار آلفا زیادتر باشد، پیش‌بینی‌های الگوریتم بیشتر به سمت گره‌های اطراف هر گره متمایل می‌شود.

روش های مبتنی بر درجه

ATTACHMENT PREFERENTIAL

- تنها از درجه گره ها برای رتبه بندی لینک ها استفاده می شود.
- احتمال اینکه یک یال جدید دارای نقطه پایانی باشد، متناسب با درجه ی گره ی است.
- اگر گراف مورد بررسی گراف نویسندگان همکار باشد، منطقی به نظر می رسد که نویسندگانی که فعال ترند (نویسندگانی که دارای همکارهای بیشتری باشند) احتمال ایجاد ارتباط های جدیدتر بیشتر دارند.
- گراف نویسندگان همکار گرافی است که در آن گره ها نویسندگان و یال ها نشان دهنده ی یک همکاری علمی دو نویسنده در نوشتن یک مقاله را نشان می دهد.

مقایسه روش های پیش بینی

- روش Katz از آن جایی که مبتنی بر مسیر می باشد، در مواردی که قطر گراف کم باشد، نتایج خوبی خواهد داشت.
- دو روش AdamicAdar و CommonNeighbor به علت استفاده از ویژگی تعداد گره های مشترک میان دو گره، تقریباً از نتایج شبیه به یکدیگر برخوردارند. هر چند که در تمامی مجموعه ها داده ای روش AdamicAdar عملکردی بهتری داشته است.
- روش PreferentialAttachment در حالتی که درجه گره ها زیاد باشد، از نتایج خوبی برخوردار است. لزوماً بالا بودن متوسط درجه گره ها در گراف باعث موفقیت این روش نمی شود. ولی می توان به طور قطع بیان نمود که اگر متوسط درجه خروجی گره ها پایین باشد این روش از نتایج خوبی برخوردار نمی باشد.
- روش AdamicAdar در مواردی که ضریب خوشه بندی بالا و چگالی گراف کم باشد، نتایج بهتری نسبت به سایر روش ها دارد.

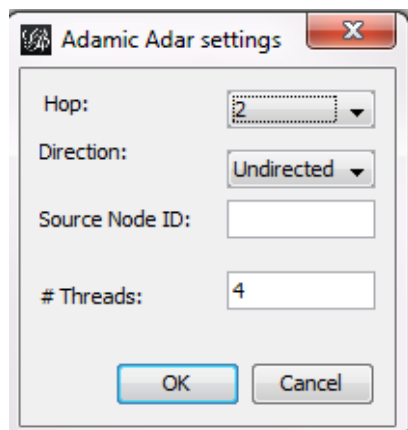
مقایسه روش های پیش بینی

- بهترین زمان اجرا متعلق به روش Preferential Attachment می باشد. روشی مقیاس پذیر محسوب می شود که با افزایش سایز شبکه، زمان اجرای آن به شدت افزایش پیدا نکند. روش هایی مثل SimRank، RootedPageRank و Katz روش هایی هستند که با افزایش سایز شبکه، زمان اجرای آن ها افزایش چشم گیری پیدا می کند. پس نمی توان از این روش ها برای گراف های بسیار بزرگ استفاده نمود.
- روش Resource Allocation در بیشتر موارد بهتر از الگوریتم Adamic Adar بوده است.

LINK PREDICTION

تنظیمات الگوریتم Adamic Adar

- با توجه به شکل می‌توان موارد زیر را مطرح نمود:
- Hop: منظور از Hop یا گام آن است که گره‌هایی با چه فاصله‌ای دورتر از گره مبدأ را باید برای پیش‌بینی لینک مورد بررسی قرار داد. برای مثال اگر مقدار گام برابر با ۲ باشد، تنها گره‌هایی با فاصله دو گام از گرهی مبدأ مورد بررسی قرار می‌گیرند. در صورتی که Hop برابر با صفر باشد، تمامی گره‌های موجود در گراف برای پیش‌بینی لینک مورد بررسی قرار می‌گیرند. لازم به ذکر است که برای محاسبه‌ی گام جهت لینک‌ها نیز مد نظر قرار می‌گیرند.
- Direction: در صورتی که گراف جهت‌دار باشد، مورد استفاده قرار می‌گیرد.



LINK PREDICTION

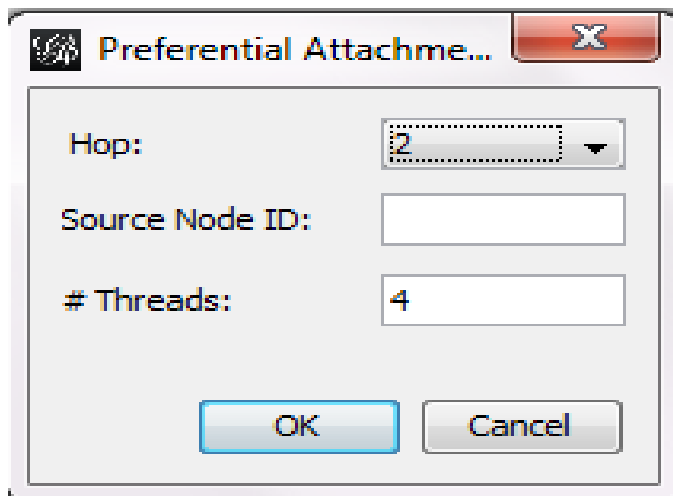
● گاهی تنها نیاز است که لینک‌های اطراف یک گره خاص مورد پیش‌بینی قرار گیرند. در این حالت با قرار دادن ID گره مربوطه در تنظیمات یک الگوریتم پیش‌بینی لینک، تنها لینک‌های اطراف آن گره مورد پیش‌بینی قرار می‌گیرند.

● **Threads #:** از آنجایی که الگوریتم‌های پیش‌بینی لینک به صورت Multi Thread اجرا می‌شوند، کاربر می‌تواند تعداد Thread هایی مورد نظر برای اجرای الگوریتم را مشخص نماید. به صورت پیش‌فرض تعداد Thread ها بر حسب تعداد هسته‌های سیستم مورد بررسی مشخص می‌شوند. برای مثال اگر یک سیستم دو هسته‌ای باشد، افزونه پیش‌بینی لینک به طور خودکار دو برابر تعداد هسته‌ها، برای الگوریتم مورد نظر Thread ایجاد می‌کند.

LINK PREDICTION

تنظیمات الگوریتم Preferential Attachment

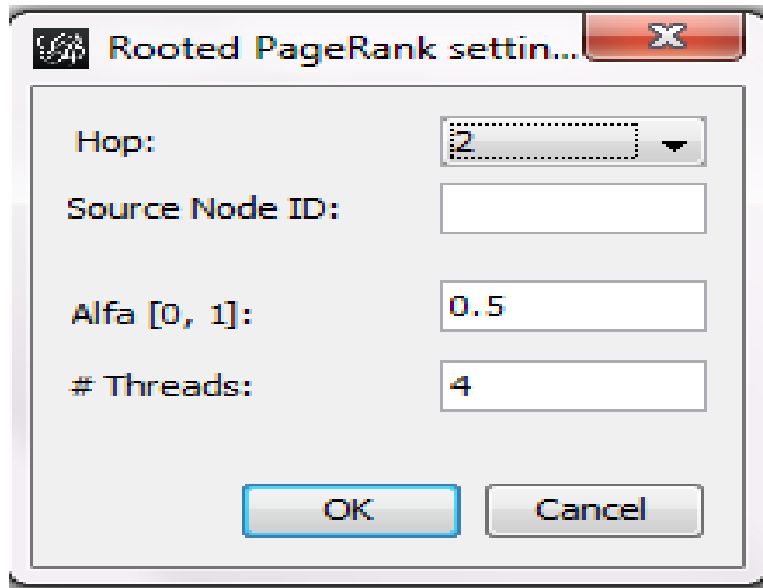
برای این الگوریتم بخش به نام Direction وجود ندارد. علت این امر ماهیت خود الگوریتم‌هاست به طوری که جهت برای پیش‌بینی لینک توسط این الگوریتم‌ها معنایی ندارد.



LINK PREDICTION

تنظیمات الگوریتم Rooted PageRank

Alfa میزان احتمال بازگشت گام بردار تصادفی به گره مبدأ است



Rooted PageRank settin...

Hop: 2

Source Node ID:

Alfa [0, 1]: 0.5

Threads: 4

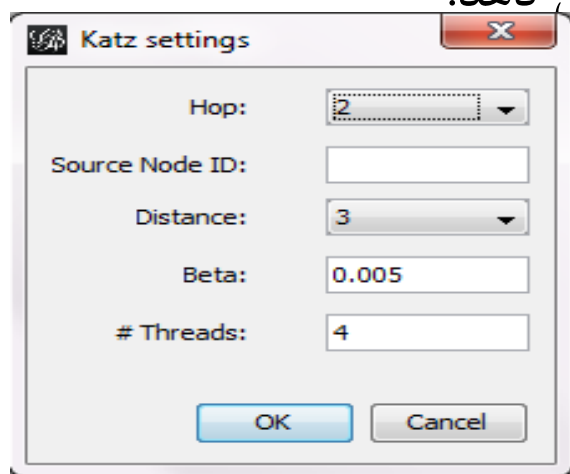
OK Cancel

LINK PREDICTION

تنظیمات الگوریتم Katz

Hop و Source Node Id و Threads همان معیارهای قبلی هستند و معیار Distance مجموعه تمامی مسیرهای به طول L از گره X به گره Y است.

معیار Beta میزان افزایش پیش بینی لینک میان گره ها در نواحی پرجمعیت نسبت به نواحی کم جمعیت را نشان می دهد.



خطای بازماندگان



SAMPLING

نمونه برداری

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

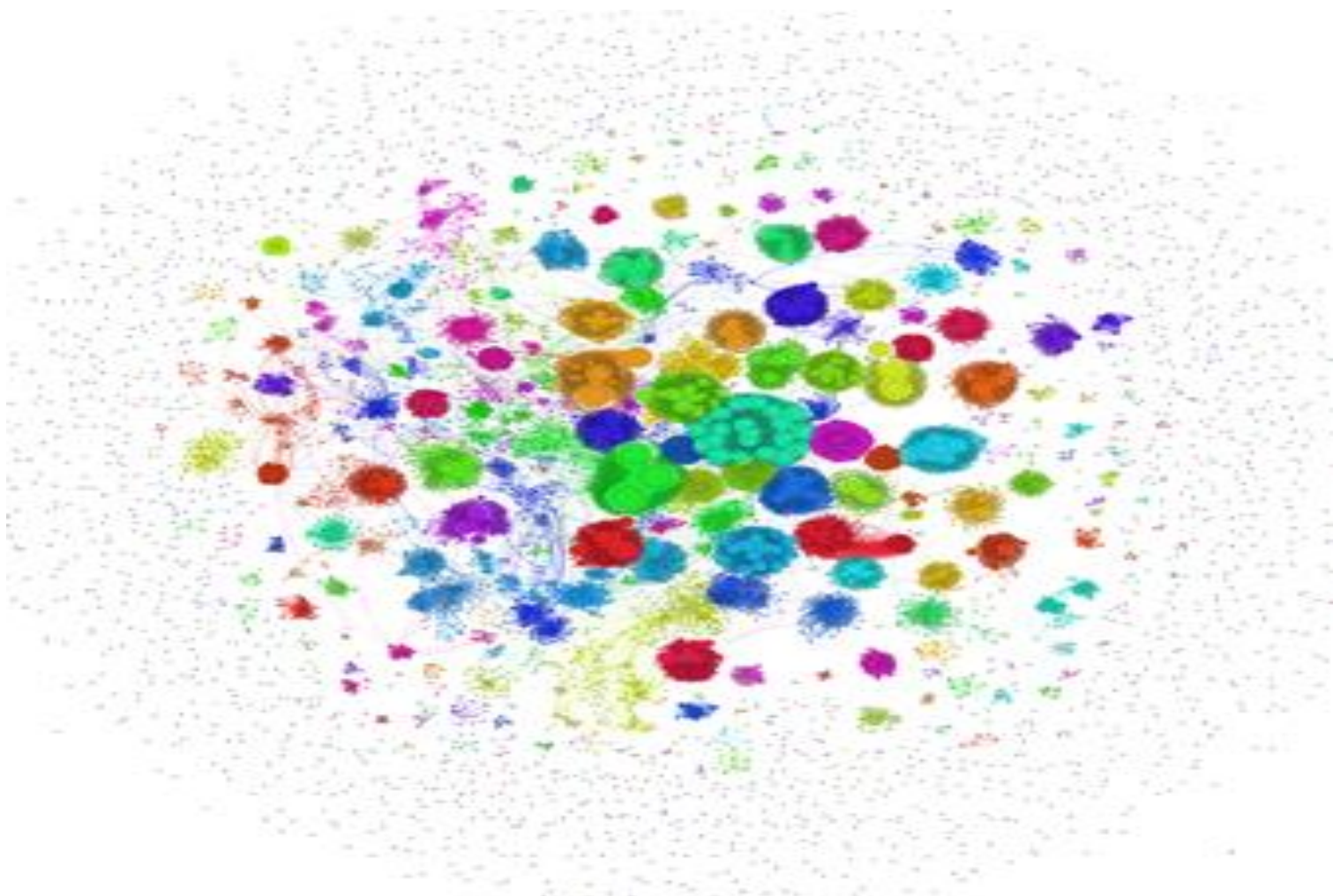
علیرضا حبیبی

نمونه برداری

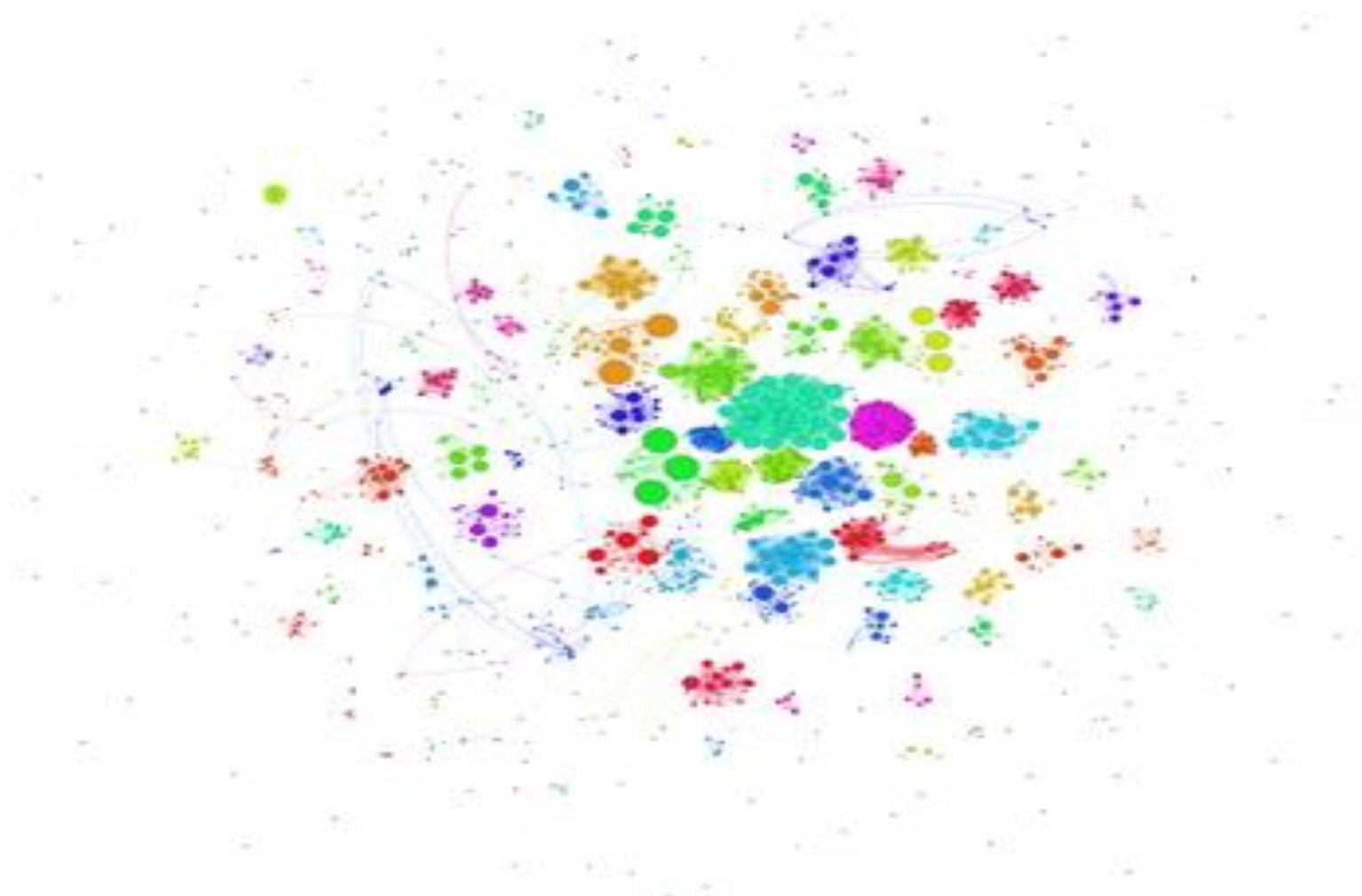
SAMPLING

- یکی از راه‌های حل مسایل پیچیده محاسباتی روی گراف‌های بزرگ، محاسبه تقریبی پارامترها است.
- این محاسبات تقریبی میتواند به صورت انجام پذیرد:
 - محاسبه تقریبی روی گراف اصلی
 - محاسبه دقیق روی گراف فرعی شبیه به گراف اصلی.
- محاسبه دقیق از نوع دوم بر روی گراف فرعی شبیه به گراف اصلی صورت می‌پذیرد که این گراف با **ساده کردن** گراف اصلی بدست می‌آید. به این روش نمونه برداری گراف می‌گویند.

گراف اصلی



گراف نمونه برداری شده



نمونه برداری مبتنی بر انتخاب تصادفی نودها

◉ Random Node:

- اولین و بدیهی ترین روش برای نمونه برداری **انتخاب تعدادی از نودها به صورت تصادفی** با احتمال انتخاب یکنواخت است.
- این الگوریتم توزیع توانی گراف را حفظ نمی کند.

◉ Page Rank:

- انتخاب نودها می تواند با احتمال غیر یکنواخت انجام شود.
- مثلاً احتمال انتخاب یک نود می تواند متناسب با رتبه صفحه آن نود باشد.

نمونه برداری مبتنی بر انتخاب تصادفی نودها

Random Degree Node

- میتوانیم نودها را با احتمالی متناسب با درجه آنها انتخاب کند.
- این روش به سمت انتخاب نودهای دارای درجه بالا سو دارد.
- این روش توزیع درجه را حفظ نخواهد کرد.
- زیرا نودهای دارای درجه بالا را که نودهای خاصی هستند انتخاب می کند.
- همین طور این گراف نمونه بسیار چگال تر از گراف اصلی خواهد بود.

نمونه برداری مبتنی بر انتخاب تصادفی یالها

Random Edge

- شبیه انتخاب تصادفی نودها می توانیم یال ها را نیز به صورت تصادفی و با توزیع یکنواخت انتخاب کنیم.
- در این روش گرافهای نمونه خلوت هستند و در نتیجه قطر بزرگی دارند.
- ساختار انجمن ها در آن ها به چشم نمی خورد.
- گراف به سمت نودهایی که دارای درجه بالایی هستند سو دارد.

نمونه برداری مبتنی بر انتخاب تصادفی یالها

Random Node-Edge

- روش دیگر ترکیب انتخاب گره و یال به صورت تصادفی است.
- گره ها را به صورت تصادفی و با توزیع یکنواخت انتخاب کنیم. سپس از میان یالهای گره ای که انتخاب کرده ایم یکی از یال هایش را برگزینیم.
- این روش بر خلاف RE به سمت نودهای دارای درجه بالا سو ندارد.

Hybrid

- میتوان RE و RNE را با هم ترکیب کرد و یک روش ترکیبی یا Hybrid بدست آورد.
- در این روش با احتمال p یک گام از روش RNE و
- با احتمال $1-p$ یک گام از روش RE را انجام می دهیم.

نمونه برداری مبتنی جستجوی تصادفی

Random Node Neighbors

- یک نود را به صورت تصادفی انتخاب می کنیم و سپس کلیه نودهای همسایه ای که آن نود به آنها خروجی دارد را انتخاب می کنیم.
- این روش برای خواندن گرافهای بزرگ مناسب است
- در این روش ساختار انجمنی گراف و توزیع درجه ورودی آن به هم می ریزد.

نمونه برداری مبتنی جستجوی تصادفی

Random Walk

- در ابتدا یک نود را به صورت تصادفی یکنواخت انتخاب می کنیم
- سپس روی گراف حرکت تصادفی انجام می دهیم.
- یعنی با احتمال c به نود شروع برمیگردیم و با احتمال $1-c$ از میان نودهای همسایه آن نود یکی را انتخاب کرده و به آن نود می رویم.

Random Jump

- تنها تفاوت این روش با روش پیمایش تصادفی این است که در پرش تصادفی با احتمال c به یک نود تصادفی منتقل می شویم (نه به نود شروع).

FOREST FIRE

این روش که آتش سوزی جنگل نامیده می شود با ایده گرفتن از تکامل گراف ها در زمان توسعه یافته است.

در این روش یک نود به صورت تصادفی انتخاب می شود و این نود آتش زده می شود. آتش با احتمالی به یال های متصل به آن سرایت کرده و از آن طریق به نودهای همسایه اش سرایت می کند. این آتش سوزی به همین صورت از نودهای همسایه آتش گرفته به نقاط دیگر به صورت احتمالی سرایت می کند.

این مدل ۲ متغیر دارد :

■ احتمال آتش سوزی رو به جلو P_f

○ احتمال سرایت آتش سوزی به نودهایی که از نودهای آتش گرفته لینک ورودی دارند.

■ احتمال آتش سوزی رو به عقب P_b

○ احتمال سرایت آتش سوزی به نودهایی که از نودهای آتش گرفته لینک خروجی دارند.

گراف اصلی قبل از نمونه برداری

۱۰۸۷۶	تعداد نودها
۳۹۹۹۴	تعداد یالها
(۱۰۰۰) ۱۰۸۷۶	تعداد نودهای بزرگترین جزء دارای همبندی ضعیف
(۱۰۰۰) ۳۹۹۹۴	تعداد یالهای بزرگترین جزء دارای همبندی ضعیف
(۰.۳۹۷) ۴۳۱۷	تعداد نودهای بزرگترین جزء قویا همبند
(۰.۴۶۹) ۱۸۷۴۲	تعداد یالهای بزرگترین جزء قویا همبند
۰.۰۰۶۲	ضریب خوشه بندی متوسط
۹۳۴	تعداد مثلثها
۰.۰۰۱۸۰۷	درصد مثلثهای بسته
۵.۴	قطر موثر (۹۰ درصد نودها)

مقایسه نتایج نمونه برداری

PR	FF	RJ	RW	RDN	RNN	H	RNE	RE	RN	گراف اصلی	روش نمونه برداری
0.005	0.007	0.005	0.006	0.006	0.005	0.01	0.003	0.002	0.002	0.006	متوسط ضریب خوشه بندی
8	9	7	9	8	8	6	6	6	8	9	قطر اصلی
0.97	0.98	0.9	0.9	0.91	0.89	0.84	0.8	0.86	0.9	1	درصد نودهای بزرگترین جزء دارای همبندی ضعیف
0.98	0.98	0.93	0.92	0.98	0.94	0.91	0.92	0.93	0.94	1	درصد یالهای بزرگترین جزء دارای همبندی ضعیف
0.44	0.45	0.4	0.4	0.43	0.4	0.36	0.33	0.37	0.4	0.4	درصد نودهای بزرگترین جزء قویا همبند
0.47	0.48	0.42	0.42	0.47	0.44	0.4	0.42	0.42	0.44	0.47	درصد یالهای بزرگترین جزء قویا همبند
50%	15%	65%	30%	60%	55%	65%	75%	70%	70%	100%	نرخ نمونه برداری

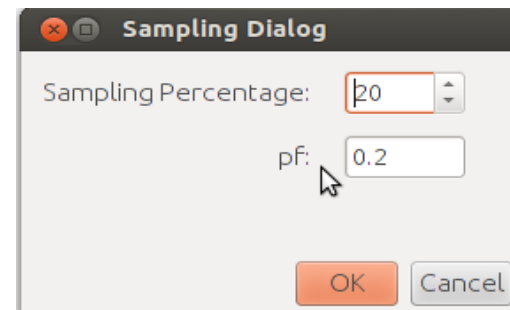
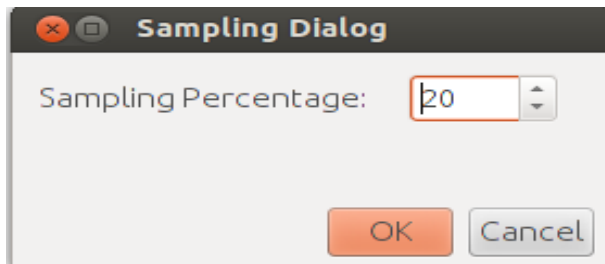
مقایسه نتایج نمونه برداری

- بر اساس نتایج آزمون روشهای مختلف بر روی مجموعه دادگان انتخاب شده چنین به نظر میرسد که روش نمونه برداری Forest Fire با وجود این که میتواند ویژگیهای گراف را حفظ کند در عین حال نیاز به نرخ نمونه برداری بالایی هم ندارد.
- پس از این روش، روشهای Random Walk و Random Jump که هر دو از روشهای مبتنی بر پیمایش تصادفی هستند بهترین نتایج را داشتند.
- از بین این دو روش Random Walk نیاز به نرخ نمونه برداری کمتری داشت و در نتیجه با این روش میتوان با گرافهای کوچکتری به نتیجه دلخواه رسید.
- ویژگی خوبی که روش Forest Fire دارد این است که نیازی به خواندن همه اطلاعات گراف در حافظه ندارد.

SAMPLING

◉ نمایی از تنظیمات الگوریتم های Sampling

در تمامی الگوریتم های نمونه برداری میزان مقدار نمونه برداری از کاربر پرسیده می شود و فقط در الگوریتم Forest Fire علاوه بر این میزان یک احتمال (Pf) نیز گرفته می شود که برابر است با احتمال آتش سوزی روبه جلو
احتمال سرایت آتش سوزی به نود های که از نود آتش گرفته لینک گرفته اند.



تحلیل های پویایی گراف

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

پیگیری انجمن‌ها

COMMUNITY TRACKING

- گراف‌های موجود در دنیای واقعی معمولاً ساختار ثابتی ندارند و در طول زمان تغییر می‌کنند.
- این تغییر می‌تواند در سطح رأس‌ها یا یال‌ها باشد.
- به عنوان مثال در یک شبکه اجتماعی واقعی، ممکن است در طول زمان یک رابطه دوستی بین دو شخص موجود در شبکه ایجاد شود، یا اینکه رابطه آن‌ها قطع شود.
- تغییر مجموعه رؤوس و یال‌های گراف در طول زمان می‌تواند بر ساختار انجمن‌های موجود در گراف نیز تأثیر بگذارد. بدین ترتیب ممکن است یک انجمن در طول زمان بزرگتر شده و شامل نودهای بیشتری شود و یا ممکن است یک انجمن با از دست دادن اعضای خودش در طول زمان کوچکتر شود.
- در طول زمان ممکن است انجمن‌های جدیدی شکل گیرد یا یک انجمن به انجمن‌های کوچکتری شکسته شود.

پیگیری انجمن‌ها

COMMUNITY TRACKING

○ پیگیری انجمن‌ها :

■ از گراف موجود در نقاط زمانی مختلف تصویر (snapshot) می‌گیرند.

○ سپس با استفاده از روش‌های یافتن انجمن، انجمن‌های هر تصویر را شناسایی می‌کنند.

○ سپس، انجمن‌های موجود در تصاویر متوالی با یکدیگر تطبیق داده می‌شوند.

○ در نهایت، این روش‌ها به منظور شناسایی نحوه تغییر انجمن‌ها در طول زمان، مجموعه‌ای از رخدادها را تعریف کرده و تغییرات انجمن‌ها را ردیابی می‌کنند.

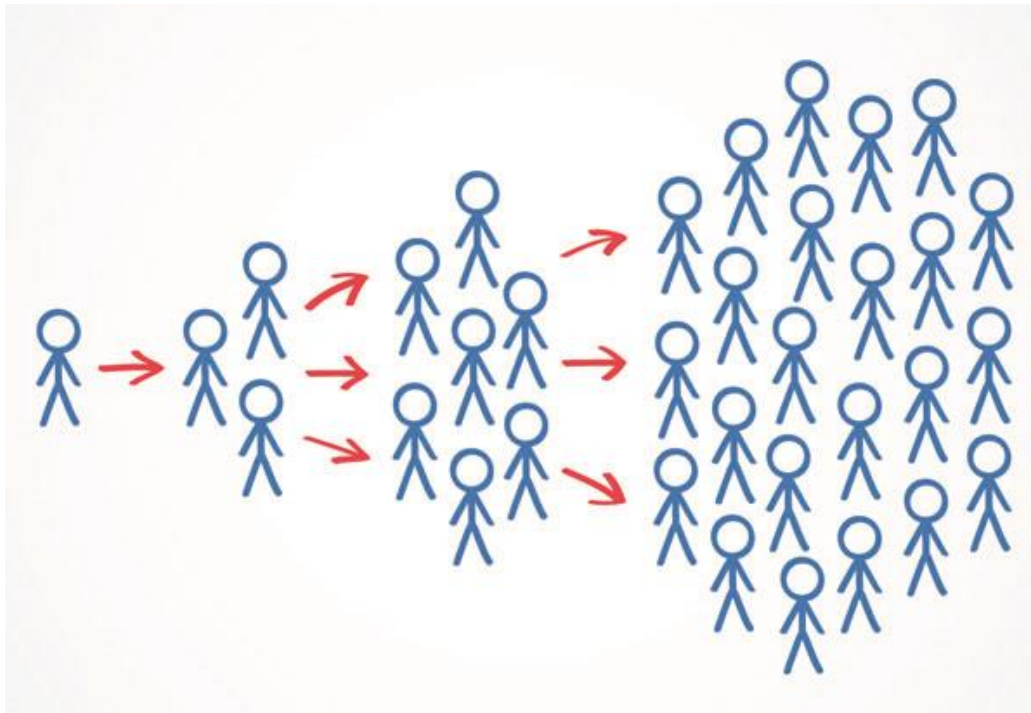
تحلیل انتشار DIFFUSION

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:
علیرضا حبیبی

◎ انتشار بیماری‌های واگیردار ویروسی و یا بیماری‌های واگیردار اجتماعی گرفته تا انتشار حوادث و خرابی‌ها، از شایعه گرفته تا گسترش یک عقیده یا دین و مذهب، از تبلیغ یک محصول گرفته تا گسترش یک فناوری، از انتشار اخبار گرفته تا انتشار رأی.



مدل های انتشار

○ روش های مبتنی بر توپولوژی

■ روش اکتشافی

■ روش حریصانه

○ مدل انتشار مستقل IC

○ مدل آستانه خطی LT

○ روش انتشار گرما HD

○ مدل اپیدمیولوژی

○ مدل انتشار فعالیت

جمع بندی

- در مدل‌های انتشار روش HD به دلیل در نظر گرفتن گره‌های غیرفعال در جریان اطلاعاتی می‌تواند بهتر از روش‌های LT و IC عمل کند.
- روش‌های گسترش یافته دیگری نیز وجود دارد که اغلب آن‌ها از روش‌های LT و IC گرفته شده‌اند. این روش‌ها همگی در شبکه‌های کوچک کاربرد دارند.
- کند بودن روش‌های حریصانه و فعالیت باعث شده تا محققان از روش‌های اکتشافی برای تحلیل شبکه استفاده کنند. از روش‌های حریصانه اغلب در ارزیابی استفاده می‌شود.

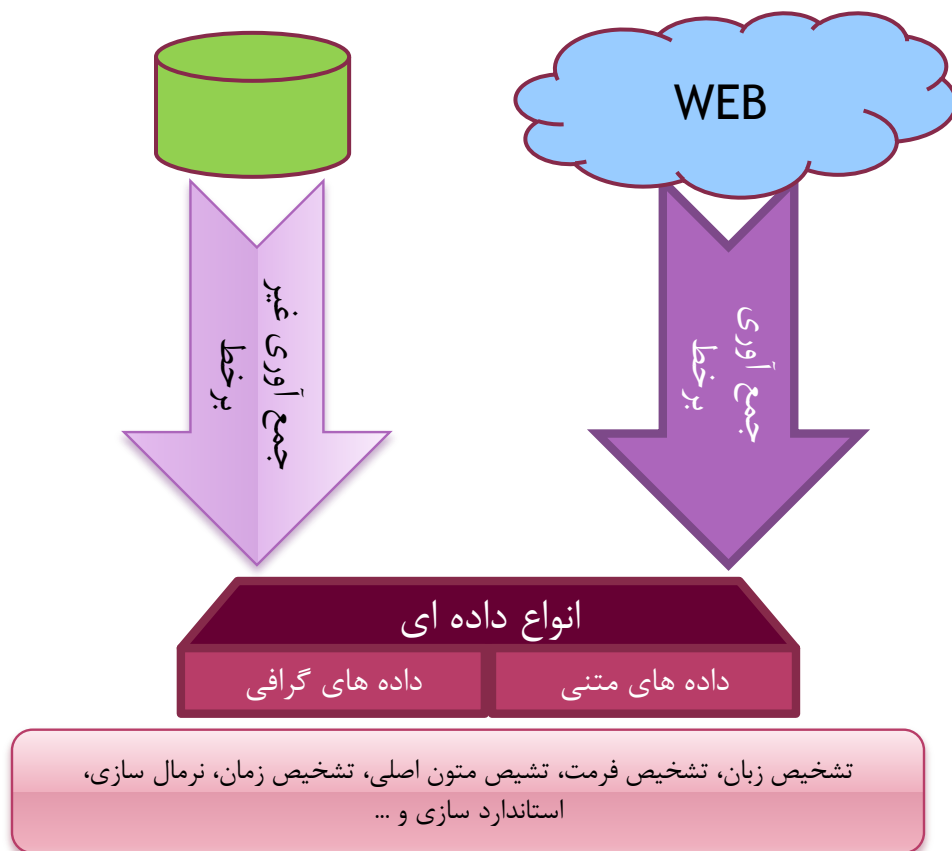
انواع روش ها جمع آوری

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:
علیرضا حبیبی

انواع روش ها جمع آوری



- خزشگر ها
- خزشگرهای کانونی
- RSS خوان ها
- وب تراش ها
- واسط های نرم افزاری
- خزشگرهای وب تاریک و وب عمیق
- خزشگرهای پروتکل شبکه ها
- خزشگرهای موبایلی
- ...

گراف های بزرگ

www.bigdata-ir.com

Telegram: [@BigData_Channel](https://t.me/BigData_Channel)

مدرس:

علیرضا حبیبی

روش های پردازش گراف های بزرگ

- روش پردازش گراف Vertex-Centrice راس محور
- روش برنامه نویسی MapReduce نگاشت / کاهش
- روش حافظه اشتراکی

- ◉ Giraph
- ◉ Graphchi
- ◉ GraphX

پایگاه داده های مبتنی بر گراف

Neo4j ●

OrientDB ●

Titan ●

خدای چنان کن سرانجام کار
تو خوشنود باشی و ما رستگار

بِسْمِ اللَّهِ الرَّحْمَنِ الرَّحِيمِ