به نام خدا

تمرین سوم شبکه‌های اجتماعی

معیارهای مرکزیت

محمد ناصری

۸۱۰۱۰۰۴۸۶

دی ۱۴۰۰

مقدمه

روش جمع آوری داده

گراف ضمیمه حاصل گرداوری اطلاعات واقعی از کانالهای تلگرامی در طول مدت یک ماه اخیر است. این کانالها در مورد شرط بندی مطلب یا مطالبی منتشر کرده اند. جهت اطلاع شما و برای اینکه تحلیل دقیقتری از نودهای گراف داشته باشید، کوئری جستجوی مطالب در کانالهای تلگرامی به صورت زیر بوده است:

- یا باید یکی از کلمات/عبارات زیر در مطلب آمده باشد (یعنی or عبارات زیر):

bet|شرطبندی|شرط بندی|مسابقه پوکر|بازی پوکر|پوکر بازی|پوکربازی|کازینو|کازینوی آنلاین|اپشن پیش بینی|ضرایب شگفت انگیز|ضرایب بالا|بازی رولت|بازی جذاب رولت|بازی انفجار|بازی جذاب انفجار|فوربت|من و تو بت|هزار بت|هزاربت|نکس بت|نکسبت|بت رایون|بترایون|ولف بت|بت فوروارد|بت فروارد|فوروارد بت|فوروارد بت|فروارد بت|بت ورزش|بت فوت|آریان بت|مل بت|کانون بت|کانن بت|وان کیک بت|وان ایکس بت|آی آر بت|ددی بت|بت وب|بت پلاس|بت90|بت پی|رومابت|روما بت|

- یا این که کلمه "بت "همراه با یکی از کلمات زیر آمده باشد (and کلمه بت همراه با orیکی از عبارات زیر):

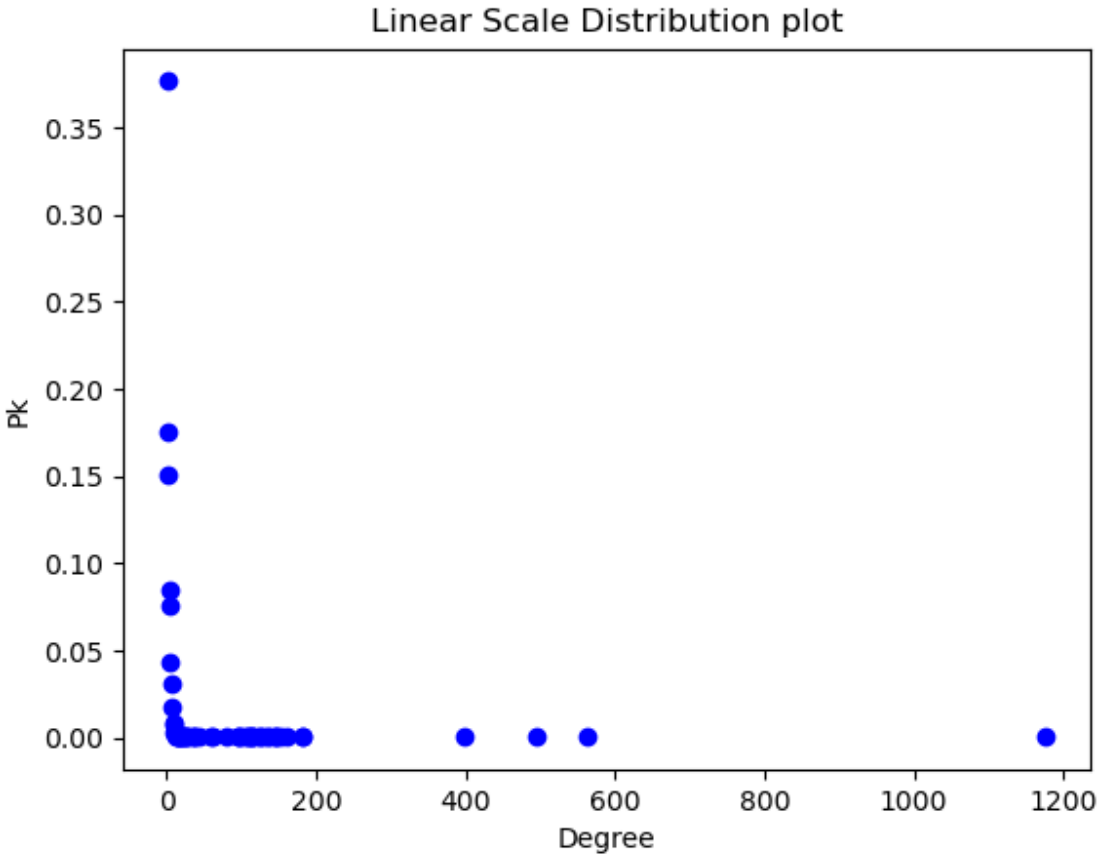
مسابقه|مسابقات|شرط|بازی|جایزه|جوایز|رولت|اسنوکر|انفجار|پوکر|اپشن|ضریب|ضرایب|لیگ|درامد سرگرمی|پرفکت مانی|برد **نکته مهمی** که باید به آن توجه بفرمایید این است که اگر مطلبی از یک فرد در کانالی فوروارد شود و این فرد در تنظیمات حریم خصوصی خود به تلگرام اجازه نداده باشد که هویت او در هنگام فوروارد برای دریافت کنندگان مطلب مشخص شود، تلگرام آن فرد را با عنوان فرستنده مخفی (HiddenSender)معرفی میکند. لذا کلیه افرادی که مطلبی از ایشان در کانالها فوروارد شده، سرجمع، با یک نود HiddenSenderدر فایل ضمیمه ذکر شده اند. لطفا در تحلیلهای خود به این نکته توجه بفرمایید. در صورتی که به این نتیجه رسیدید که حضور این نود باعث اشتباه و بهم خوردن تحلیلها میشود میتوانید آن را از جمع نودها حذف کنید

سوال اول

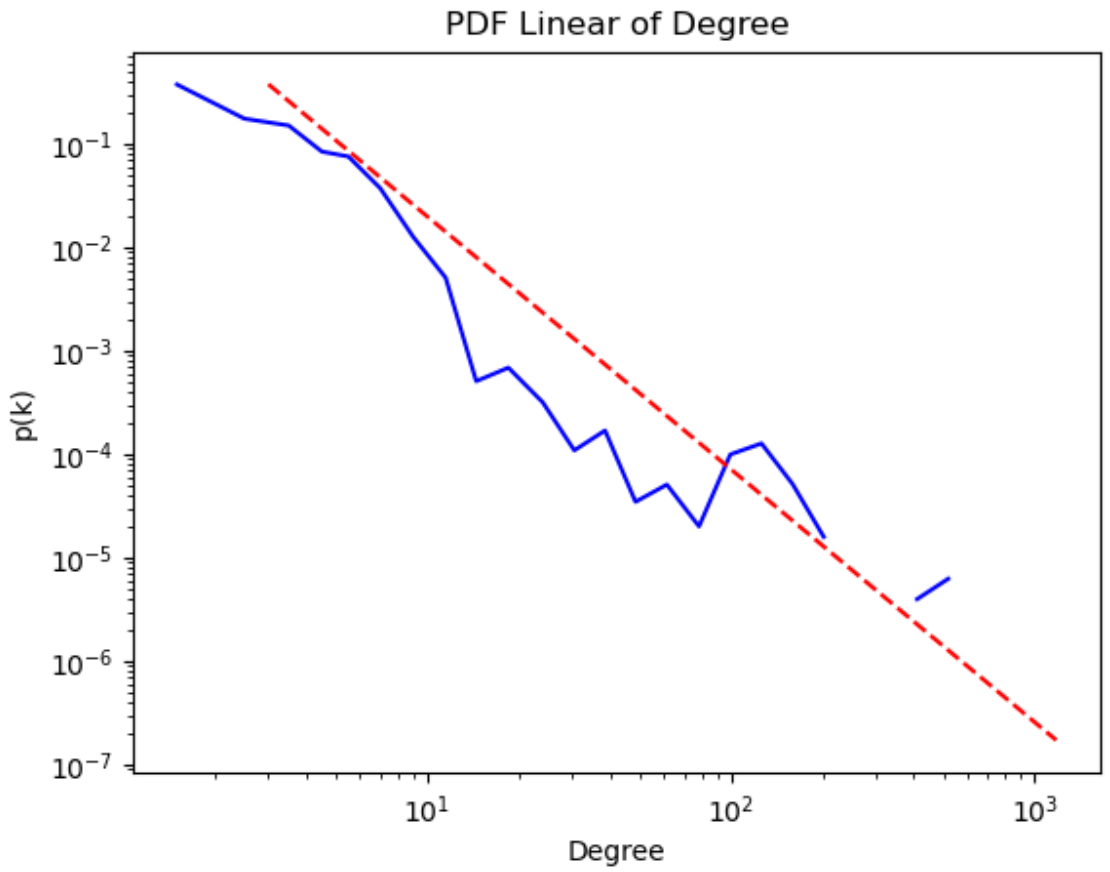
توپولوژی گراف را بررسی کنید و با رسم نمودار مشخص کنید توزیع درجه نودها از چه توزیعی تبعیت میکند. شما باید پارامترهای توزیع را نیز به صورت دقیق محاسبه کنید. همین کار را برای توزیع تعداد مطالب و جمع بازدید کانالها نیز انجام دهید.

ترسیم توزیع درجه، بخشی جدایی ناپذیر از تجزیه و تحلیل خصوصیات یک شبکه است. فرآیند با به دست آوردن N_k، تعداد گره های با درجه k شروع می شود. این را می توان با اندازه گیری مستقیم یا با یک مدل ارائه کرد. از $p_k = N_k /N$ را محاسبه می کنیم. سوال این است که چگونه p_k را رسم کنیم تا خواص آن را به بهترین نحو استخراج کنیم.

نمودار اولیه p_k برای دنباله درجات به صورت زیر است:

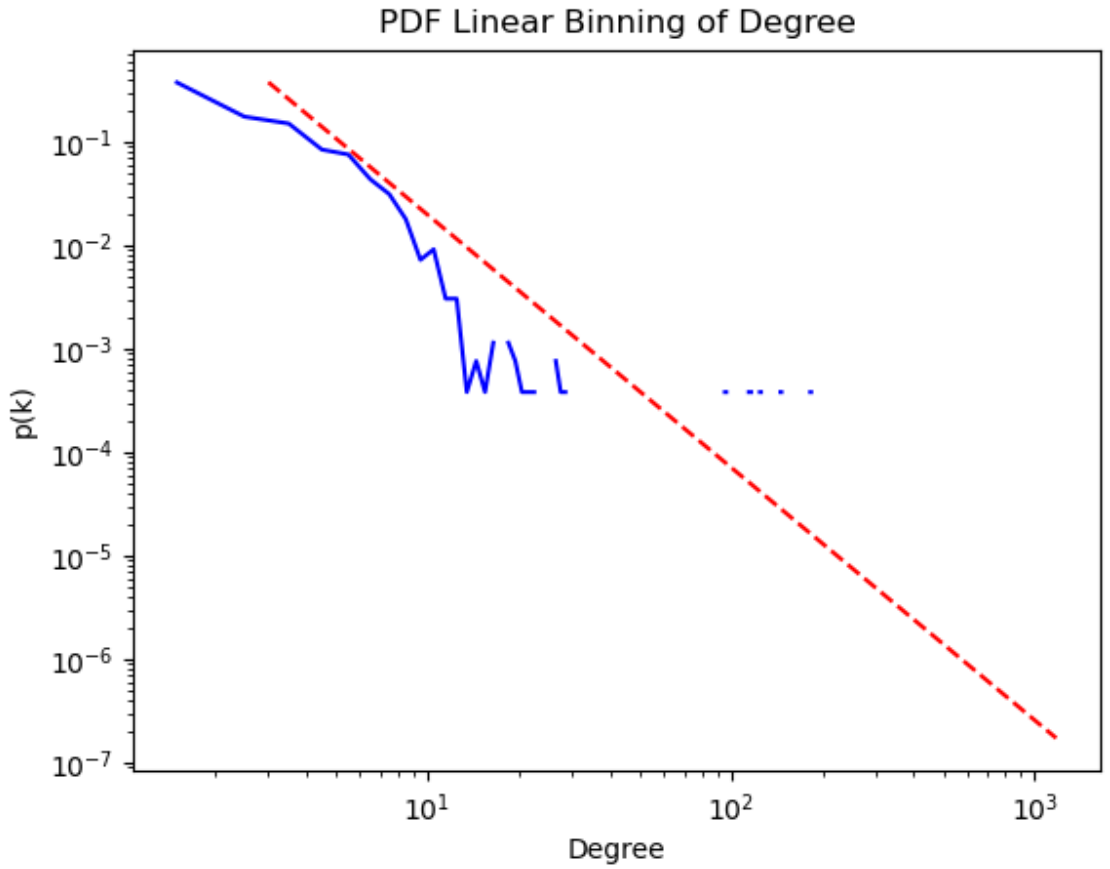


در یک شبکه بدون مقیاس، گره های متعددی با یک یا دو پیوند با چند هاب همزیستی دارند که نشان دهنده گره هایی با هزاران یا حتی میلیون ها پیوند هستند. استفاده از یک محور k خطی، گره‌های درجه کوچک متعدد در ناحیه k کوچک را فشرده کرده و آنها را نامرئی می‌کند. به همین دلیل برای نمایش بهتر این مقادیر از نمودار Log-Log استفاده میکنیم.



ناقص ترین روش (که اغلب در مقالات دیده می شود) این است که به سادگی $p_k = N_k/N$ را بر روی نمودار log-log رسم میکنند. به این حالت Bining خطی می گویند، زیرا هر سطل دارای اندازه یکسانی $\Delta k = 1$ است. برای یک شبکه بدون مقیاس، باینینگ خطی منجر به یک فلات قابل تشخیص در k بزرگ می شود که از نقاط داده متعددی تشکیل شده است که یک خط افقی را تشکیل می دهند. این فلات توضیح ساده ای دارد: معمولاً ما فقط یک کپی از هر گره درجه بالا داریم، بنابراین در ناحیه high-k یا $N_k=0$ (بدون گره با درجه k) یا $N_k=1$ (یک گره منفرد با درجه) داریم. در نتیجه باینینگ خطی یا $p_k=0$ را ارائه می‌کند، که در نمودار log-log نشان داده نمی‌شود، یا $p_k = 1/N$، که برای همه هاب‌ها اعمال می‌شود، و یک فلات در $p_k = 1/N$ ایجاد می‌کند. این فلات توانایی ما برای تخمین درجه توان را تحت تاثیر قرار میدهد.

باینینگ لگاریتمی نمونه برداری غیر یکنواخت از باینینگ خطی را تصحیح می کند. برای log-binning اجازه می دهیم اندازه bin با درجه افزایش یابد، مطمئن شویم که هر bin دارای تعداد گره های قابل مقایسه است. به عنوان مثال، می‌توانیم اندازه‌های bin را مضرب 2 انتخاب کنیم، به‌طوری‌که bin اول دارای اندازه $b_0=1$ است که شامل همه گره‌های k=1 می‌شود. دومی دارای اندازه $b_1=2$ است که شامل گره هایی با درجه k=2, 3 است. سطل سوم دارای اندازه $b_2=4$ است که شامل گره هایی با درجات k = 4,5,6,7 است. با القاء، bin_n دارای اندازه 2^{n-1} است و شامل تمام گره های با درجات $k=2^{n-1}$،2^{n+1} -1 است. ...، 2^n. توجه داشته باشید که اندازه bin می تواند با مقادیر دلخواه افزایش یابد، $b_n = c^n$، که در آن c> 1 است. توزیع درجه با $p_{\langle kn \rangle}=N_n/bn$ داده می شود، که در آن N_n تعداد گره های موجود در bin_n اندازه b_n است. و ⟨kn⟩ میانگین درجه گره ها در b_n bin است.



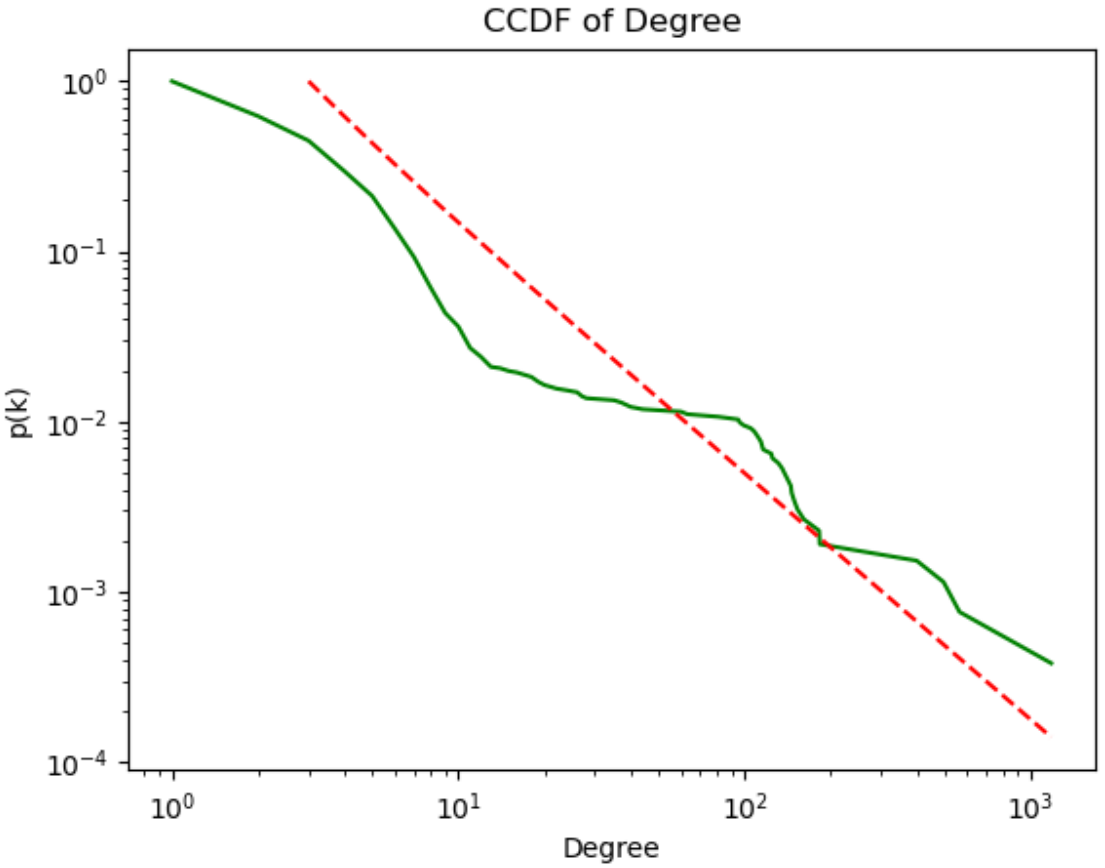
روش دیگر برای استخراج اطلاعات از p_k ترسیم توزیع تجمعی مکمل است.

$$P_k = \sum_{q=k+1}^{\infty} P_q$$

که دوباره اهمیت آماری منطقه درجه بالا را افزایش می دهد. اگر p_k از قانون توان پیروی کند، توزیع تجمعی به صورت زیر مقیاس می شود:

$$P_k \sim k^{-\gamma+1}$$

توزیع تجمعی دوباره فلات مشاهده شده برای باینینگ خطی را حذف می کند و منجر به یک ناحیه مقیاس بندی گسترده می شود که امکان برآورد دقیق تری از درجه توان را فراهم می کند.



برای پارامترهای توزیع، با کمک کتابخانه powerlaw به محاسبه پارامترهای تابع fit میپردازیم که نتایج حاصل به شرح زیر هستند:

likelihood	**5.251607478976815**
γ	**2.444815061621069**
k_{min}	**3.0**
k_{max}	**None**
D	**0.07679686858035473**

این کتابخانه تمامی فرمولهای کتاب Barabasi را پیاده‌سازی کرده و مقادیر بدست آمده از همان روش‌های کتاب میباشند. مقدار γ همان مقدار درجه توان در قانون توان میباشد که داریم

$$P_k = k^{-\gamma}$$

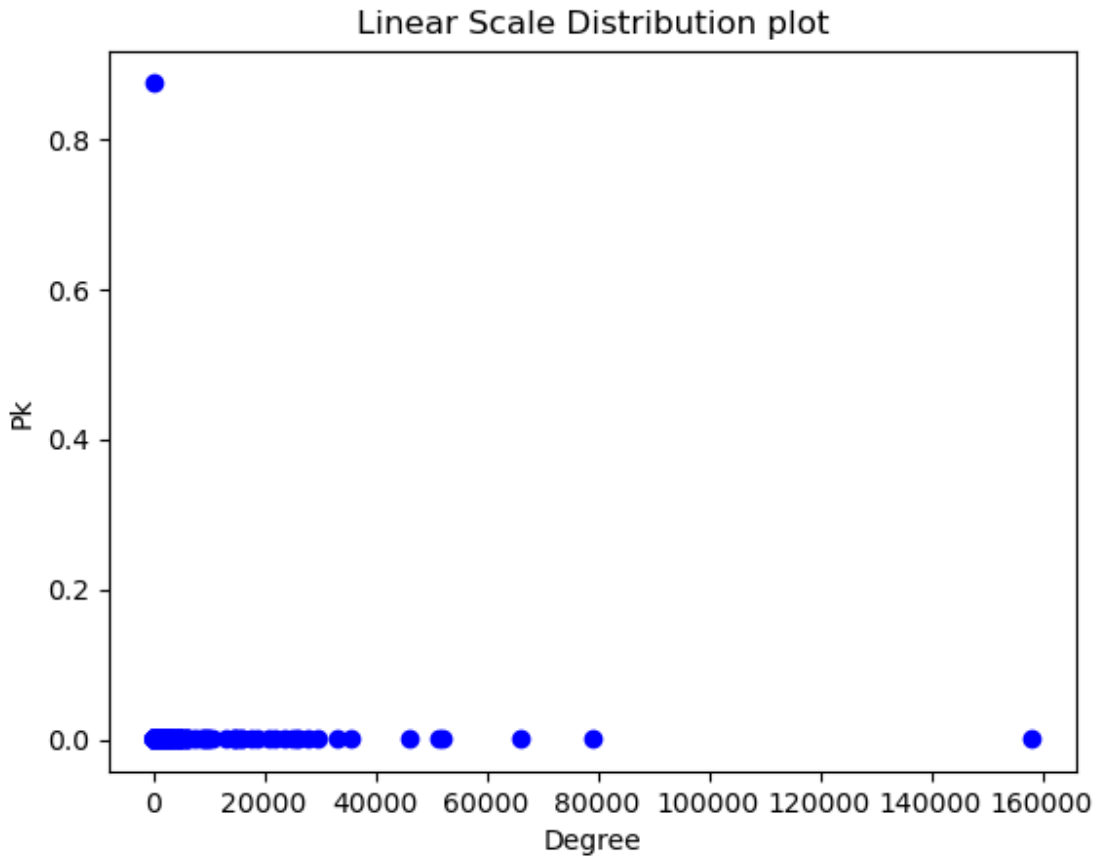
مقدار K_{min} و K_{max} نشان‌دهنده بازه Scaling برای تابع فیت هستند که در fitting procedure محسابه میشوند و مقدار likelihood مقایسه میان دو توزیع نمایی و powerlaw میباشد که با مثبت بودن نشان از powerlaw و با منفی بودن نشان دهنده توزیع نمایی میباشد.

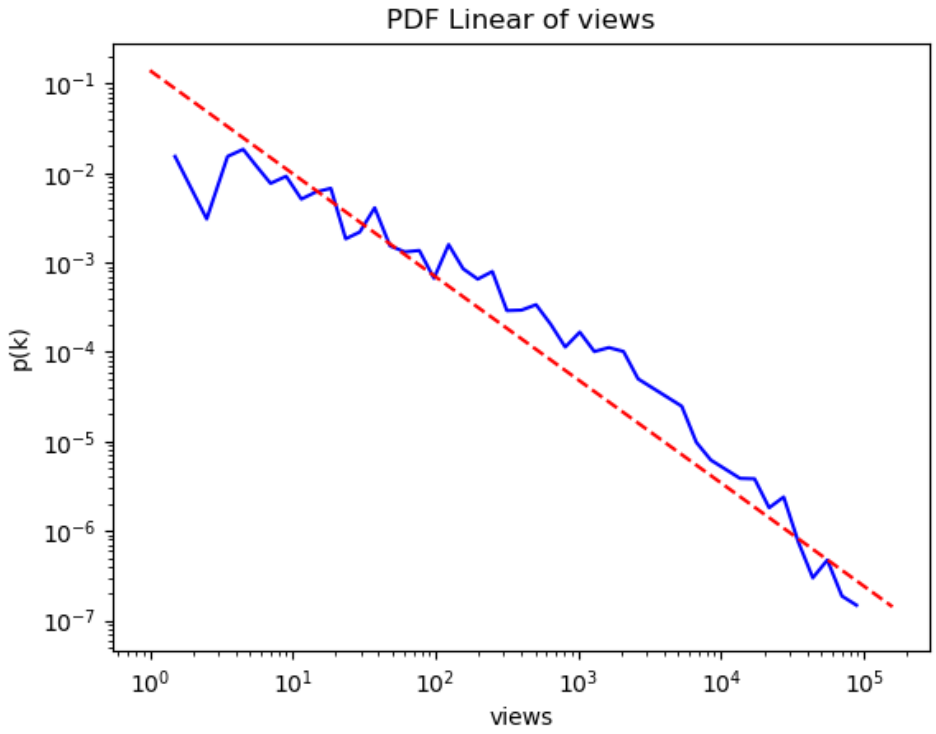
توضیحات بالا را برای count و view مجدد محاسبه میکنیم. برای هر ۳ نمودار برای اطمینان حاصل کردن بیشتر از نتایج از کتابخانه Fitter برای مطمئن شدن استفاده میکنیم که نتایج برای همه توزیع نمایی پیش بینی میشود و با مقایسه توزیع نمایی و powerlaw توسط کتابخانه Powerlaw متوجه میشویم که توزیع‌های ما شباهت بیشتری به powerlaw دارند پس از همین توزیع استفاده میکنیم. نتایج fitter برای درجات بعنوان مثال در زیر آورده شده و باقی نیز مانند همین مثال هستند.

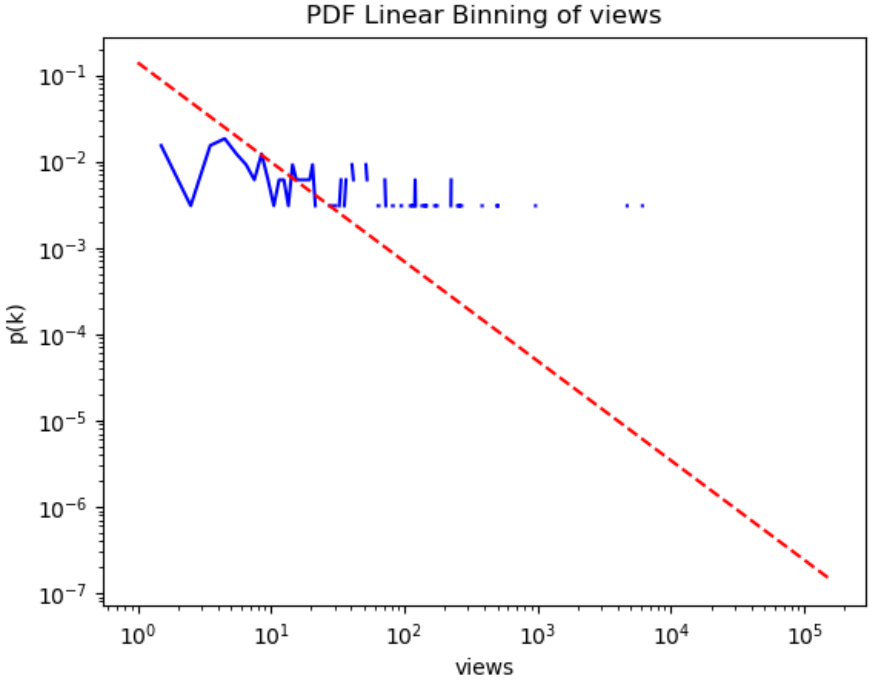
```
{'expon': {'loc': 1.0, 'scale': 4.196483180428134}}
          sumsquare_error           aic           bic  kl_div
expon            0.000612  28314.314479 -39925.041686     inf
exponpow         0.004049   4454.390968 -34975.072361     inf
chi2             0.004086   3136.313696 -34951.209811     inf
rayleigh         0.004548  48522.774682 -34679.113339     inf
cauchy           0.005000   2650.371181 -34431.092589     inf
```

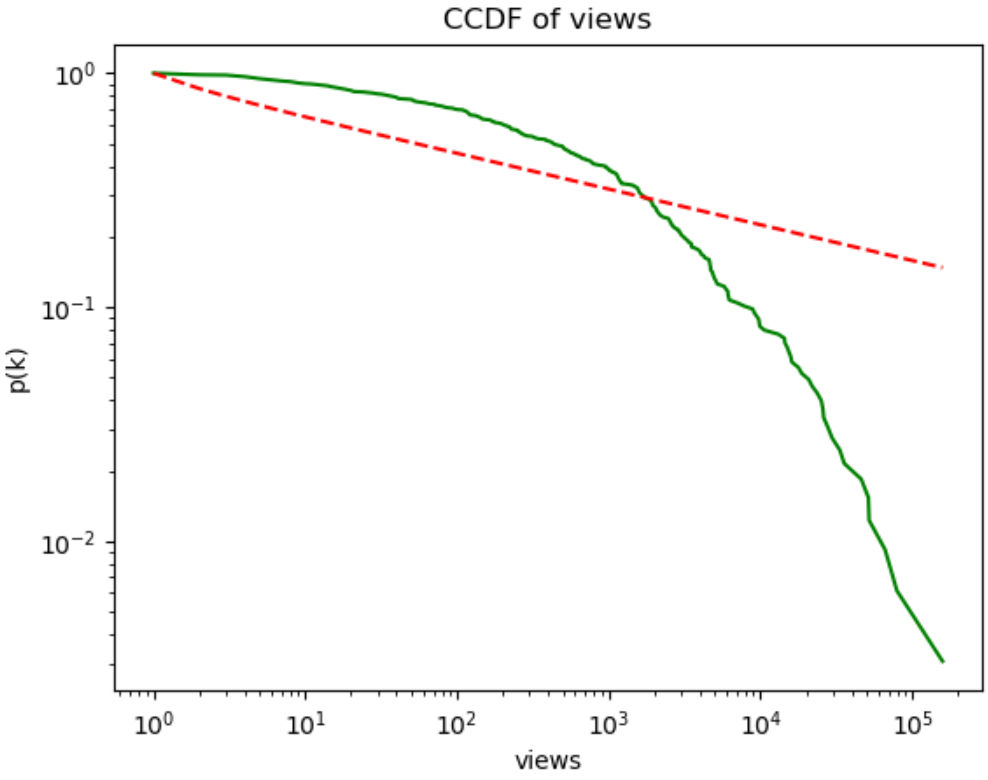
VIEWS

ابتدا با تابع مقایسه توزیع‌های کتابخانه powerlaw بررسی میکنیم که آیا تابع بدست آمده از قانون توان پیروی میکند یا نزدیک به exponential میباشد که با توجه به نتایج likelihood بدست آمده تابع به powerlaw نزدیک تر میباشد. توزیع نمایی حداقل نامزد مطلق جایگزین برای ارزیابی سنگین بودن توزیع است. دلیل آن تعریفی است: تعریف کمی معمولی heavy tail این است که به صورت نمایی محدود نشده است.




PDF Linear of views
p(k)
views
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}


PDF Linear Binning of views
p(k)
views
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{0}
10^{1}
10^{2}
10^{3}
10^{4}
10^{5}

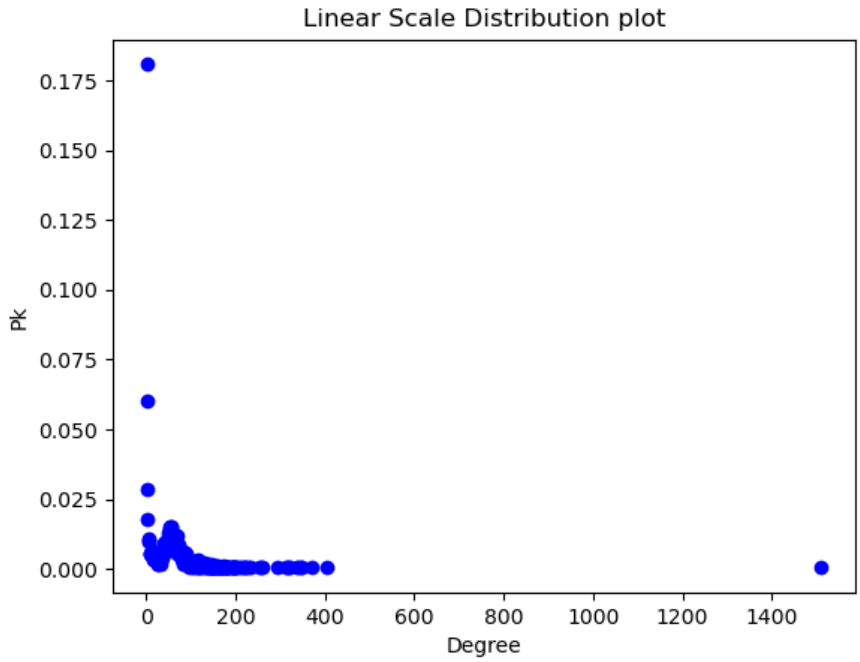


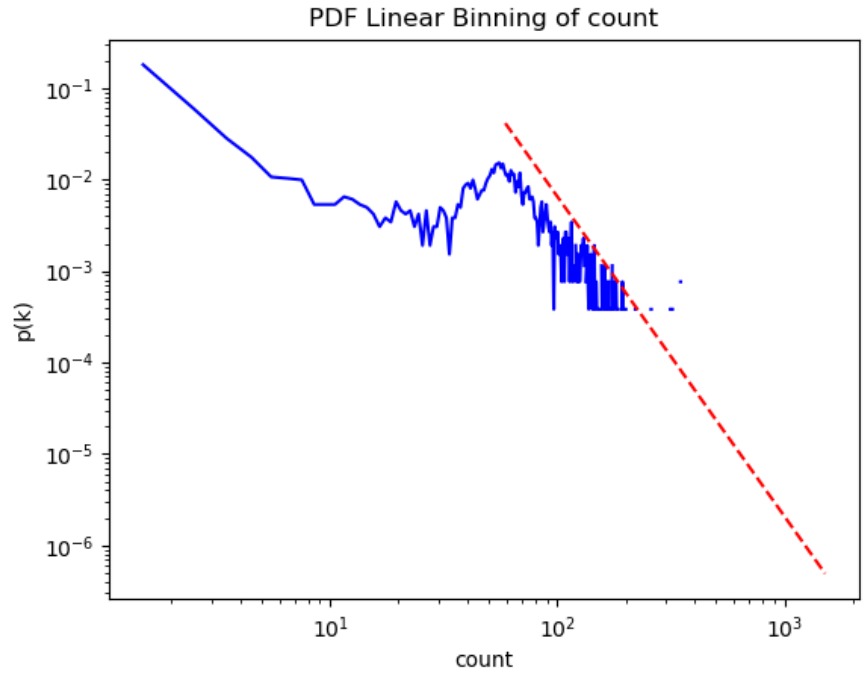
likelihood	**2.5955926988270077**
γ	**1.7728512571986585**
k_{min}	**1453.0**
k_{max}	**None**
D	**0.07027027276112507**

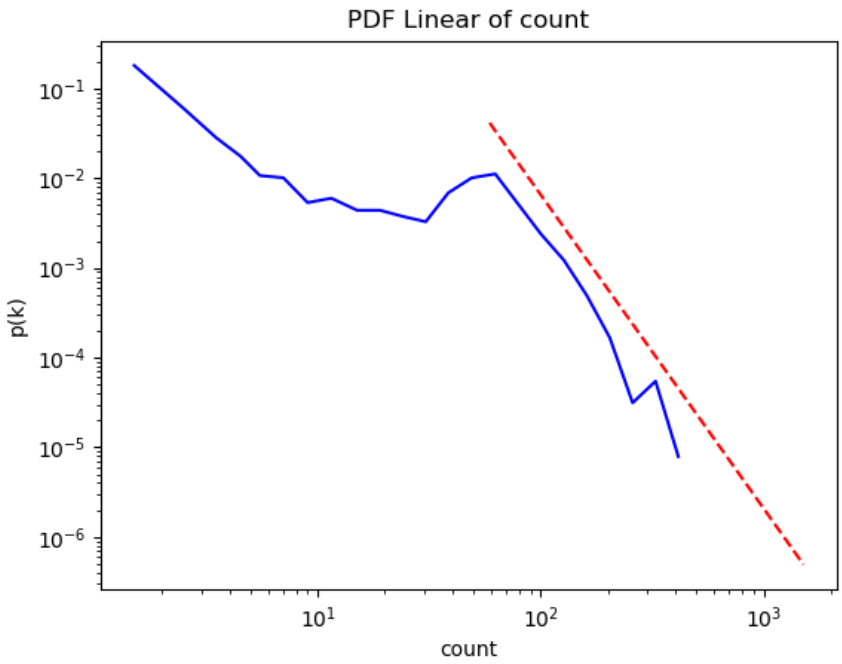
```
{'expon': {'loc': 0.0, 'scale': 473.7454128440367}}
          sumsquare_error           aic           bic  kl_div
expon        4.424965e-08  34635.718625 -64868.438796     inf
chi2         2.180579e-07   3951.729207 -60688.276453     inf
exponpow     2.241114e-07   3066.212595 -60616.643592     inf
rayleigh     2.491260e-07  41723.336260 -60347.699689     inf
norm         2.823796e-07  41987.637952 -60019.931859     inf
```

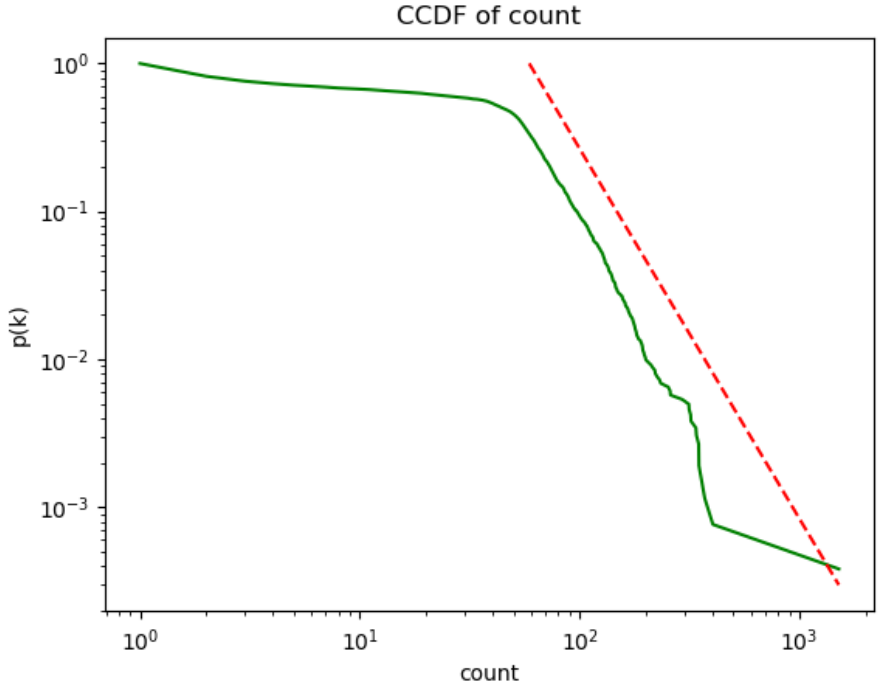
COUNT

این تابع نسبت به توابع قبلی به دلیل مقادیر نامتناسب با Powerlaw پیروی کمتری از شکل این تابع دارد.






PDF Linear of count
p(k)
count


CCDF of count
p(k)
count

likelihood	**1.6806887867781177**
γ	**3.497173294245536**
k_{min}	**59.0**
k_{max}	**None**
D	**0.024004236000087165**

```
{'expon': {'loc': 1.0, 'scale': 45.17737003058104}}
          sumsquare_error           aic           bic  kl_div
expon            0.000171   4108.500631 -43266.086594     inf
exponpow         0.000227   6625.560395 -42508.710551     inf
lognorm          0.000253   2792.200388 -42224.785560     inf
rayleigh         0.000334  18009.516918 -41512.290044     inf
norm             0.000385  24168.879077 -41135.000993     inf
```

روشی که در بالا توضیح داده شد ممکن است این تصور را ایجاد کند که تعیین توان درجه یک فرآیند پیچیده اما ساده است. در واقع این روش‌های برازش دارای محدودیت‌های شناخته شده‌ای هستند:

یک قانون توان خالص یک توزیع ایده آل است که فقط در مدل های ساده ظاهر می شود. در واقعیت، طیف وسیعی از فرآیندها به توپولوژی شبکه‌های واقعی کمک می‌کنند و بر شکل دقیق توزیع درجه تأثیر می‌گذارند. اگر p_k از قانون توان خالص پیروی نکند، روش های شرح داده شده در بالا که برای تطبیق قانون توان با داده ها طراحی شده اند، ناگزیر در تشخیص اهمیت آماری شکست خواهند خورد. در حالی که این یافته می تواند به این معنی باشد که شبکه بدون مقیاس نیست، اغلب به این معنی است که ما هنوز به درک درستی از شکل دقیق توزیع مدرک دست نیافته ایم. از این رو ما شکل عملکردی اشتباه pk را به مجموعه داده fit می کنیم.

ابزارهای آماری مورد استفاده در بالا برای آزمایش مناسب بودن Fit بر معیارهای Kolmogorov-Smirnov تکیه می کنند که حداکثر فاصله بین مدل fit شده و مجموعه داده را اندازه گیری می کند. اگر تقریباً همه نقاط داده از قانون توان کامل پیروی کنند، اما یک نقطه به دلایلی از منحنی منحرف شود، اهمیت آماری برازش را از دست خواهیم داد. در سیستم های واقعی دلایل متعددی برای چنین انحرافات محلی وجود دارد که تأثیر کمی بر رفتار کلی سیستم دارد. با این حال، حذف این "غیرطبیعی" بودن می تواند به عنوان دستکاری داده ها تلقی شود. با این حال، اگر حفظ شود، نمی توان اهمیت آماری تناسب قانون توان را تشخیص داد.

به طور خلاصه، تخمین توان درجه هنوز یک علم دقیق نیست. ما همچنان فاقد روش هایی هستیم که اهمیت آماری را به روشی که برای یک متخصص قابل قبول باشد برآورد کند. کاربرد کورکورانه ابزارهایی که در بالا توضیح داده شد اغلب منجر به تناسب هایی می شود که آشکارا روندهای داده ها را نشان نمی دهد یا به رد نادرست فرضیه قانون توان منجر می شود.

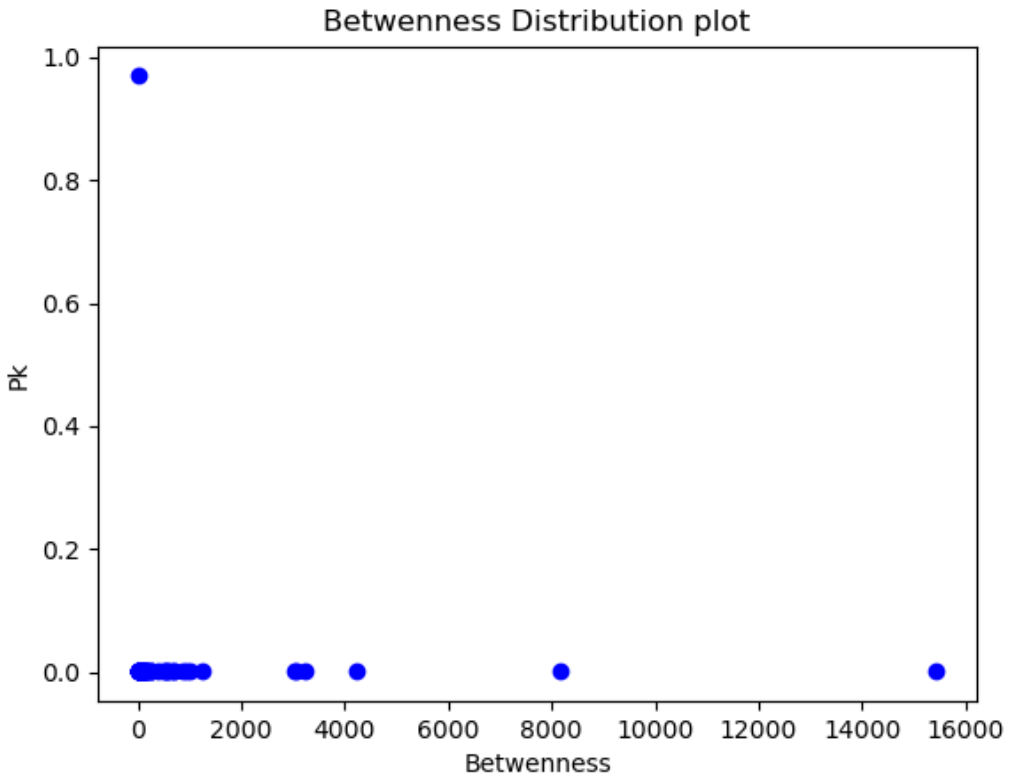
ظاهر متفاوت نمودارهای سوال یک به دلایل ذکر شده در بالا هستند زیرا مقادیر آورده شده بطور کامل از powerlaw پیروی نمیکنند.

سوال دوم

برای این کار شما باید معیارهای مرکزیت را با استفاده از ابزارهای موجود در Gephiیا با کمک گرفتن از ابزارهای دیگر محاسبه کرده و 10نود برتر هر معیار را بیابید. سپس تفسیر کنید دلیل برتر بودن آنها در آن معیار چه بوده است. توجه داشته باشید که ممکن است برای تفسیر بهتر لازم باشد محتوای مطالب این کانالها را نیز بررسی نمایید. معیارهای مرکزیتی که باید حساب کنید:

بینابینی (betweenness)

برای محاسبه مقادیر بینابینی در Gephi از محاسبات Network Diameter استفاده میکنیم. پراکندگی مقادیر بینابینی به صورت زیر است:



در مرحله بعد ۱۰ مورد با بالاترین بینابینی را بررسی میکنیم. این ده کانال عبارتند از:

1. Tanzgaradagh
2. Perpolis_persian
3. Betflood
4. S3ximovie
5. mame_twitterr
6. mitingg
7. x2betir
8. betcart
9. Takseda_music
10. footballclub

مرکزیت بینابینی برای تشخیص میزان تأثیر یک گره بر جریان اطلاعات در یک گراف است. اغلب برای یافتن گره هایی استفاده می شود که به عنوان یک پل از یک قسمت از یک گراف به قسمت دیگر عمل می کنند.

این الگوریتم کوتاه ترین مسیرهای بدون وزن را بین تمام جفت گره ها در یک نمودار محاسبه می کند. هر گره بر اساس تعداد کوتاه ترین مسیرهایی که از گره می گذرد امتیازی دریافت می کند. گره‌هایی که بیشتر در کوتاه‌ترین مسیرها بین گره‌های دیگر قرار می‌گیرند، امتیازات مرکزیت بین‌گرایی بالاتری خواهند داشت.

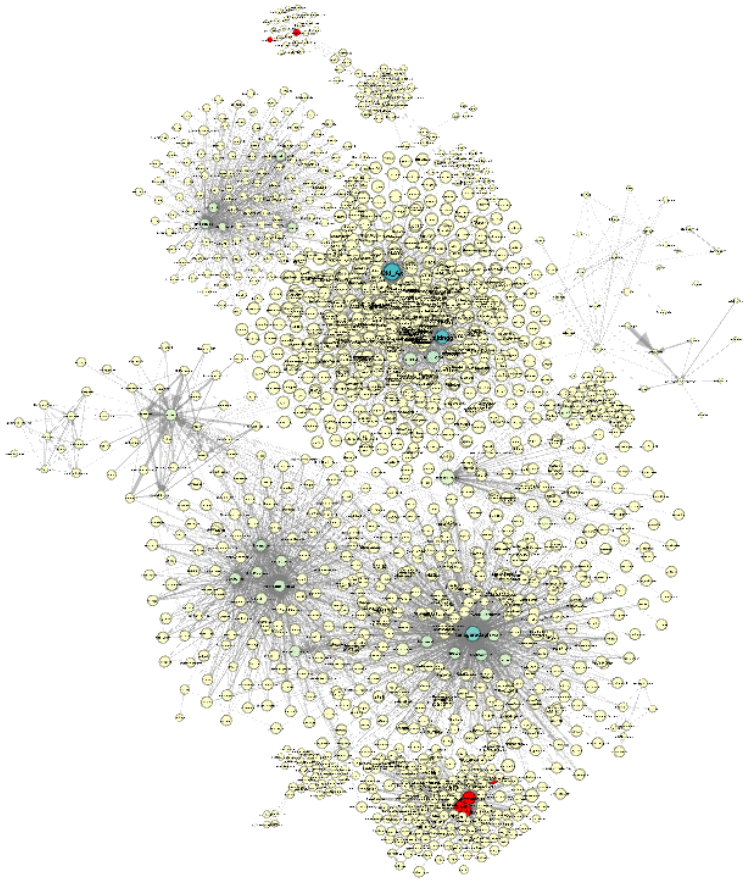
ابتدا به مفهوم بینابینی در این شبکه میپردازیم. در این شبکه جهت یالها از کانال فورواردکننده به سمت کانال مرجع یعنی نویسنده مطلب است. از طرفی میدانیم بینابینی به معنای تعداد کوتاهترین مسیرهای بین گره‌هاست است که گره انتخابی در این مسیر قرار دارد.

با توجه به تعاریف گفته شده در بالا میتوان بینابینی در این شبکه را به این صورت توصیف کرد که یک کانال برای دسترسی به اطلاعات و پیام‌های کانال دیگر در صورت عدم دسترسی مستقیم از کدام یک از دیگر کانال‌ها میتواند برای یافتن این اطلاعات استفاده کند. در واقع یعنی اگر فرض کنیم یک کاربر به دنبال یک کانال مشخص باشد و از پست‌های یک کانال شروع کند و به پیام‌های فوروارد شده رجوع کند تا به کانال بعدی برسد احتمال اینکه به کانال‌های با betweenness بالا برسد بسیار بالاست و از طریق همین کانال‌هاست که میتواند به کانال‌های دیگر دسترسی داشته باشد. به صورت کلی جریان اطلاعات بیشتر از این کانالها عبور میکند و گلوگاه ارتباطی برای کانالها هستند.

می‌توان این نتیجه‌گیری را داشت که برای تبلیغ و معرفی کانال‌های جدید هم بهتر است از این کانال‌ها استفاده کرد زیرا که احتمال اینکه کاربران جدید به آنها رجوع داشته باشند بالا خواهد بود.

نزدیکی (closeness)

مرکزیت نزدیکی روشی برای تشخیص گره هایی است که قادر به انتشار بسیار موثر اطلاعات از طریق گراف هستند.



مرکزیت نزدیکی یک گره، دوری متوسط آن (فاصله معکوس) را با سایر گره ها اندازه گیری می کند. گره هایی با امتیاز نزدیکی بالا کمترین فاصله را با سایر گره ها دارند.

در مقایسه این معیار با ۲ مشکل روبرو هستیم:

- نودها و کامپوننت‌های ایزوله و غیر قابل دسترس
- تعریف جهت یالهای گراف

مورد اول با اعمال فیلتر قابل حل میباشد. فیلترهای مورد استفاده عبارت اند از Giant Component, Partition Count . با اعمال این فیلترها تنها کامپوننت‌های متصل بزرگ باقی میمانند.

مورد دوم بدلیل نحوه تعریف جهت یالها به وجود می‌آید زیرا جهت یالهای گراف ما از فوروارد کننده به مرجع میباشد و با این تعریف از نتایج حاصل از مرکزیت نزدیکی مفهوم مناسبی نمیتوان استخراج کرد. بنابراین برای

بررسی دقیق‌تر ارتباطات کانال‌ها بر اساس این معیار، ارتباطات و یالها را بدون جهت در نظر میگیریم. با این اوصاف نتایج برتر عبارتند از

1. Old_Ax
2. Tanzgaradagh
3. Takseda_music
4. Mitingg
5. pok_Movie
6. mitingg
7. motor_gazii
8. ProfileSet1
9. Shol_Quiz
10. shol_english

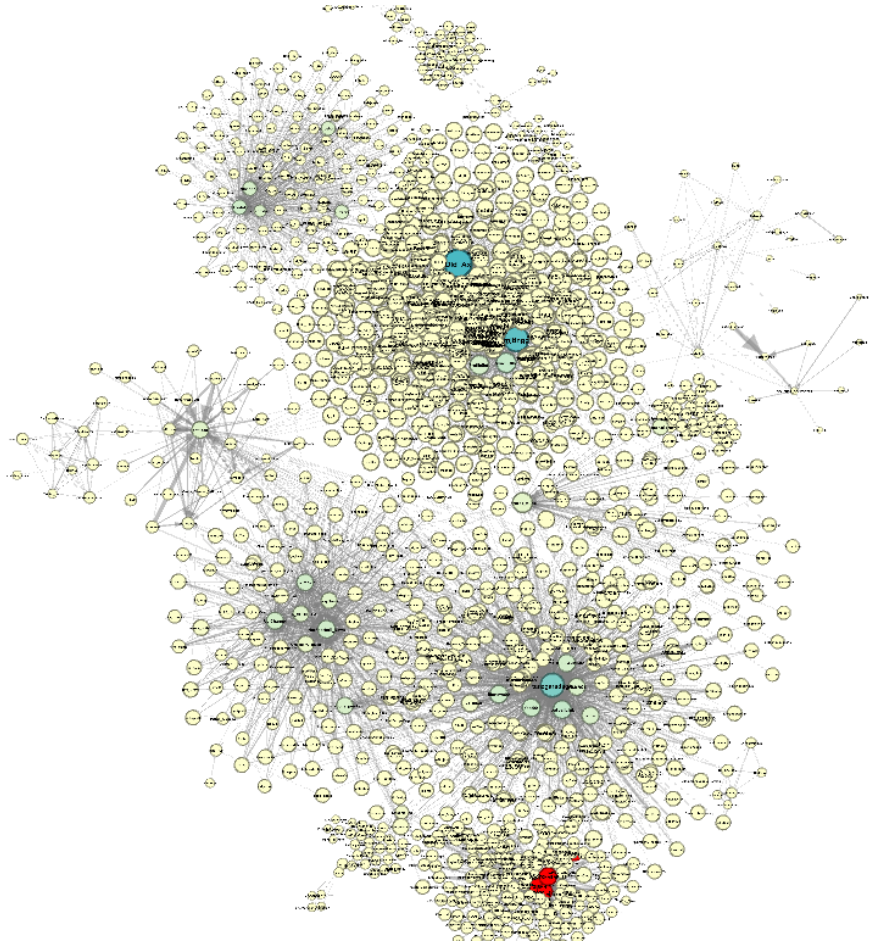
این کانال‌ها از لحاظ interaction داشتن با کانالهای دیگر مرکزی نزدیک به کانالها هستند و میتوان از آنها برای نشر داده و اطلاعات استفاده کرد.

کارایی (efficiency)

مرکزیت هارمونیک (همچنین به عنوان مرکزیت ارزشی شناخته می شود) نوعی از مرکزیت نزدیکی است که برای حل مشکلی که فرمول اصلی هنگام برخورد با گراف های غیرمتصل داشت ابداع شد. همانند بسیاری از الگوریتم های مرکزیت، از حوزه تحلیل شبکه های اجتماعی سرچشمه می گیرد. مرکزیت هارمونیک توسط Marchiori و Latora در Harmony in the Small World در حالی که تلاش می‌کردند مفهوم معقولی از «متوسط کوتاه‌ترین مسیر» ارائه کنند، پیشنهاد شد.

آنها روش متفاوتی را برای محاسبه میانگین فاصله با فاصله مورد استفاده در الگوریتم Closeness Centrality پیشنهاد کردند. الگوریتم مرکزیت هارمونیک به جای جمع کردن فواصل یک گره با تمام گره های دیگر، معکوس آن فواصل را جمع می کند. این باعث می شود که با مقادیر نامتناهی مقابله کند

مرکزیت هارمونیک به عنوان جایگزینی برای مرکزیت نزدیکی پیشنهاد شد و بنابراین موارد استفاده مشابهی دارد.



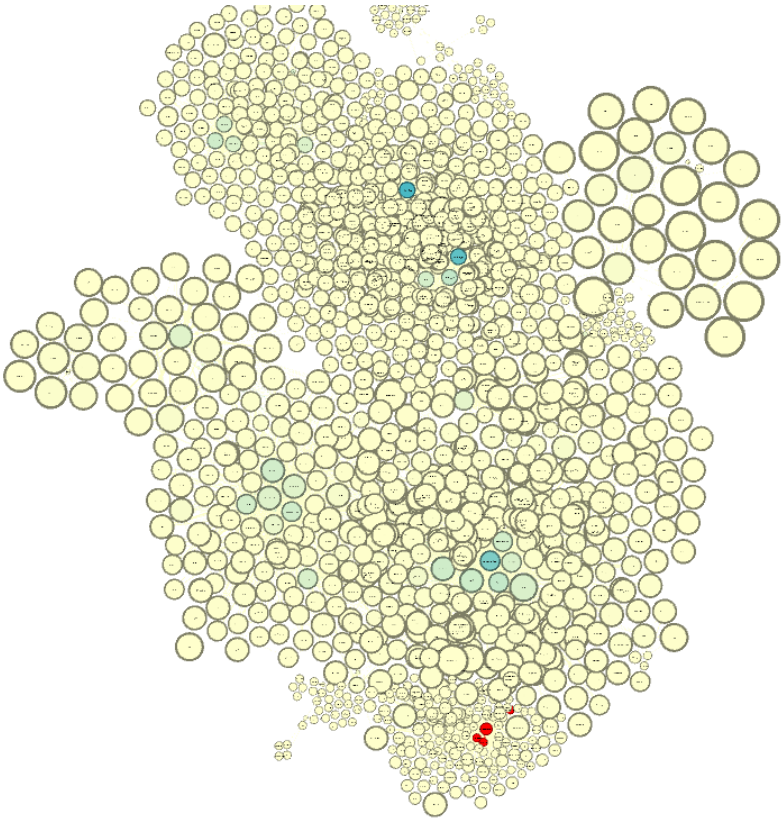
به‌عنوان مثال، اگر می‌خواهیم مکانی در شهر را برای قرار دادن یک سرویس عمومی جدید شناسایی کنیم تا به راحتی برای ساکنان قابل دسترسی باشد، ممکن است از آن استفاده کنیم. اگر می‌خواهیم پیامی را در رسانه‌های اجتماعی منتشر کنیم، می‌توانیم از این الگوریتم برای یافتن تأثیرگذارهای کلیدی استفاده کنیم که می‌توانند به ما در رسیدن به هدفمان کمک کنند.

1. Old_Ax
2. Mitingg
3. Tanzgaradagh
4. Takseda_music
5. Mind_C0ld
6. pok_Movie
7. ProfileSet1
8. Footballclub
9. fifanews90
10. foot90

این کانال‌ها از لحاظ interaction داشتن با کانالهای دیگر مرکزی نزدیک به کانالها هستند و میتوان از آنها برای نشر داده و اطلاعات استفاده کرد و میتوان مانند معیار نزدیکی از نتایج برداشت کرد.

نامرکزیت (eccentricity)

به عنوان حداکثر فاصله یک راس از راس دیگر تعریف می شود. حداکثر فاصله بین یک راس تا همه راس های دیگر به عنوان نامرکزیت راس در نظر گرفته می شود که با e(V) نشان داده می شود. حداکثر مقدار برابر قطر گراف است و حداقل مقدار این معیار برای گراف را شعاع گراف می گویند.



از آنجایی که برای بدست آوردن نودهایی که از لحاظ معیار نامرکزیت بهتر هستند یعنی طولانی ترین فاصله کمتری دارند باید نتایج گفی را برعکس کنیم، از مینیمم ها استفاده میکنیم و برترها به شرح زیر هستند:

1. Tanzgaradagh
2. pok_Movie
3. Footballclub
4. fifanews90
5. foot90
6. YASHILKANDIMIZ
7. FOOTKHABAR_85
8. motor_gazii
9. betcart
10. Shol_Quiz

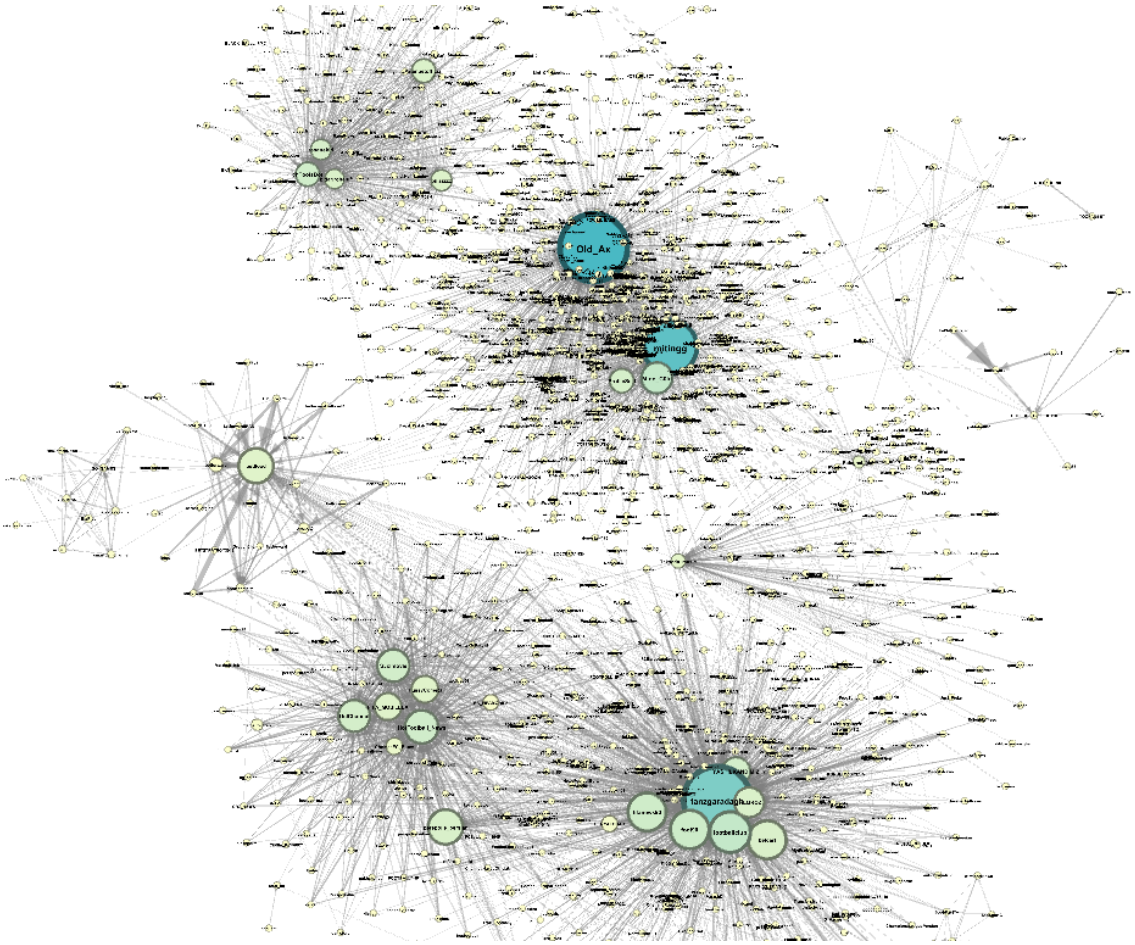
بردار ویژه (Eigenvector)

Eigenvector Centrality الگوریتمی است که تأثیر گذاری گره ها را اندازه گیری می کند. روابطی که از گره‌های با امتیاز بالا سرچشمه می‌گیرند، بیشتر به امتیاز یک گره کمک می‌کنند تا ارتباطات از گره‌های با امتیاز پایین. امتیاز بردار ویژه بالا به این معنی است که یک گره به گره های زیادی متصل است که خودشان امتیاز بالایی دارند.

الگوریتم بردار ویژه مقدار مرتبط با بزرگترین مقدار ویژه مطلق را محاسبه می کند. برای محاسبه آن مقدار ویژه، الگوریتم رویکرد تکرار توان را اعمال می کند. در هر تکرار، امتیاز مرکزیت برای هر گره از امتیازات همسایگان ورودی آن مشتق می شود. در روش تکرار توان، بردار ویژه پس از هر تکرار L2-نرمال می شود.

هنگام استفاده از الگوریتم مرکزیت بردار ویژه باید به نکاتی توجه داشت:

- امتیاز مرکزی برای گره هایی که هیچ رابطه ورودی ندارند به 0 همگرا می شوند.
- به دلیل نرمال سازی درجات گره های از دست رفته، گره های درجه بالا تأثیر بسیار زیادی بر امتیاز همسایگان خود دارند.



رتبه بندی صفحه گوگل و مرکزیت کاتز انواعی از مرکزیت بردار ویژه هستند.

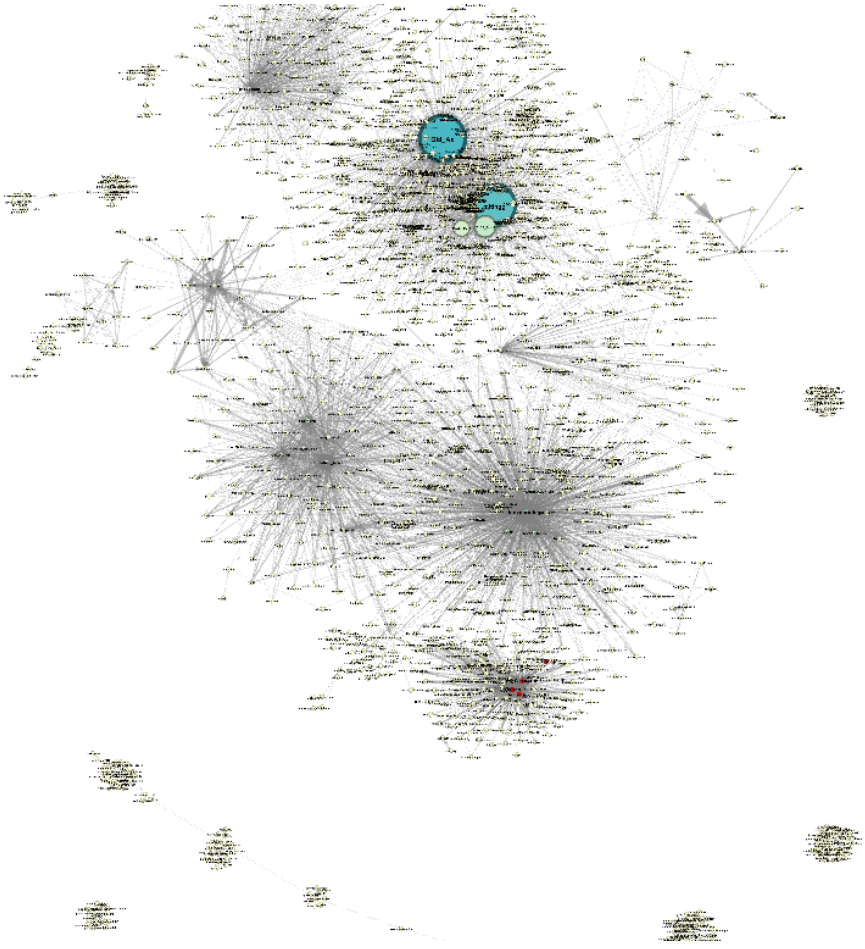
1. Old_Ax
2. Tanzgaradagh
3. Mitingg
4. Footballclub
5. foot90
6. fifanews90
7. betcart
8. Betflood
9. perspolis_persian
10. HotFootball_News

HITS

جستجوی مبحث ناشی از هایپرلینک (HITS) یک الگوریتم تجزیه و تحلیل پیوند است که گره ها را بر اساس دو امتیاز، یک امتیاز مرکز و یک امتیاز اعتبار، رتبه بندی می کند. امتیاز اعتبار، اهمیت گره را در شبکه تخمین می زند. امتیاز هاب ارزش روابط آن را با گره های دیگر تخمین می زند.

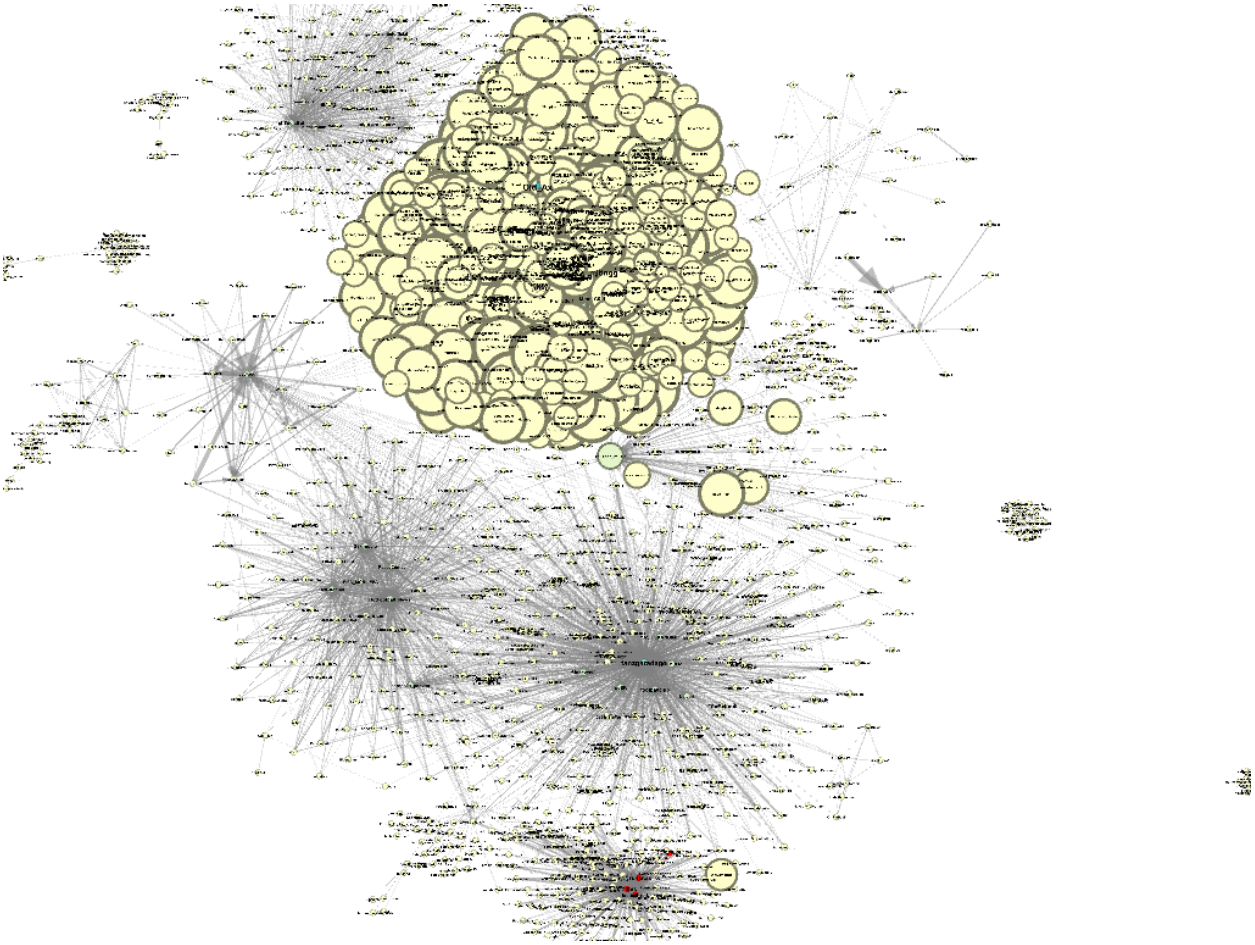
معتبرترین کانال‌ها عبارتند از:

1. Old_Ax
2. Mitingg
3. Mind_C0ld
4. ProfileSet1
5. Refaghat_ghat
6. Takseda_music
7. Tanzgaradagh
8. cinema_filmmmm
9. footballclub
10. foot90



برترین هاب‌ها عبارتند از:

1. Arz_onliine
2. Youtube_org
3. Beest_wishess
4. Marasli_irrr
5. Shabhaypataya
6. Pe_Mesle_Paeiz
7. Nationalgeographicfarsi
8. Evill
9. ShockTube
10. Whoman88

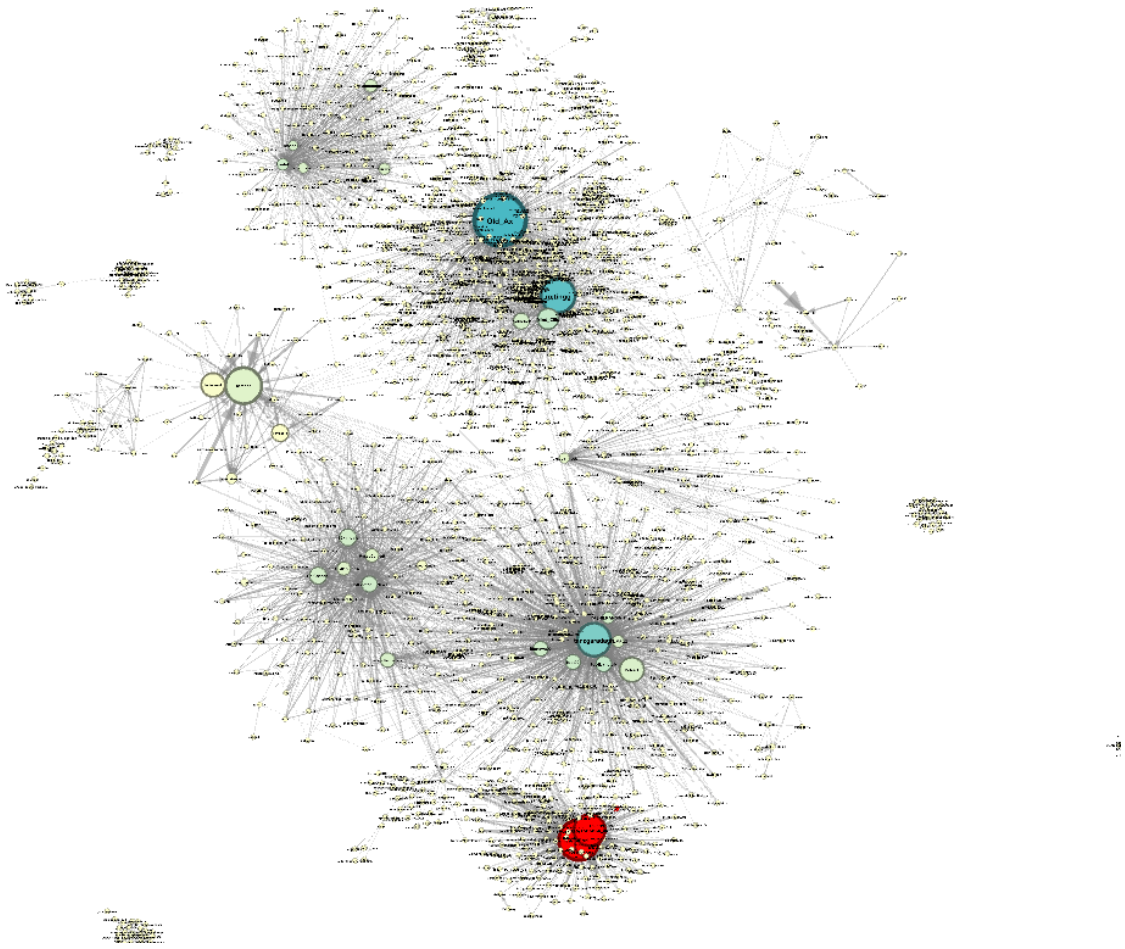


ایده پشت Hubs و Authorities از بینش خاصی در ایجاد صفحات وب در زمانی که اینترنت در ابتدا در حال شکل گیری بود، سرچشمه می گرفت. یعنی صفحات وب خاصی که به عنوان هاب شناخته می شوند، به عنوان دایرکتوری های بزرگی عمل می کردند که در واقع در اطلاعاتی که نگهداری می کردند معتبر نبودند، اما به عنوان مجموعه ای از فهرست گسترده ای از اطلاعات استفاده می شدند که کاربران را به صفحات معتبر دیگر هدایت می کرد. به عبارت دیگر، یک مرکز خوب نمایانگر صفحه ای است که به بسیاری از صفحات دیگر اشاره می کند، در

حالی که یک مرجع خوب نشان دهنده صفحه ای است که توسط هاب های مختلف لینک شده است. به صورت کلی HITS از هاب ها و اعتبارها برای تعریف رابطه بازگشتی بین صفحات وب استفاده می کند. با توجه به تعاریف جهت یال‌ها که برای این گراف داریم میتوان تعریف هاب را اینگونه تفسیر کرد که کانال‌هایی هستند که بیشتر از باقی پیام از کانال‌های دیگر فوروارد کرده و بیشتر از بقیه از مطالب کانال‌های دیگر استفاده میکنند. همچنین برای تفسیر مراجع میتوان آنها را کانال‌هایی در نظر گرفت که مطالب را بین دیگر کانال‌ها پخش و توزیع میکنند و باقی کانال‌ها مطالب را از این کانال‌ها فوروارد میکنند.

رتبه صفحه (page rank)

الگوریتم PageRank بر اساس تعداد روابط ورودی و اهمیت گره های منبع مربوطه، اهمیت هر گره را در نمودار اندازه گیری می کند. فرض اساسی به طور کلی این است که یک صفحه فقط به اندازه صفحاتی که به آن پیوند دارند اهمیت دارد. رتبه صفحه ریاضی برای یک شبکه ساده به صورت درصد بیان می شود. (Google از یک مقیاس لگاریتمی استفاده می کند.) صفحه C رتبه صفحه بالاتری نسبت به صفحه E دارد، حتی اگر پیوندهای کمتری به C وجود داشته باشد زمانیکه یک پیوند به C از یک صفحه مهم می آید و از این رو ارزش بالایی دارد. PageRank یک الگوریتم تجزیه و تحلیل پیوند است و یک وزن عددی را به هر عنصر از یک مجموعه اسناد hyperlink شده، مانند شبکه جهانی وب، با هدف "اندازه گیری" اهمیت نسبی آن در مجموعه اختصاص می دهد. این الگوریتم ممکن است برای هر مجموعه ای از موجودیت ها با نقل قول ها و مراجع متقابل اعمال شود.



هنگام استفاده از الگوریتم PageRank باید به نکاتی توجه داشت:

- اگر هیچ رابطه ای از داخل یک گروه از صفحات به خارج از گروه وجود نداشته باشد، گروه به عنوان یک تله عنکبوت در نظر گرفته می شود.
- هنگامی که شبکه ای از صفحات در حال تشکیل یک چرخه بی نهایت هستند، کاهش رتبه ممکن است رخ دهد.
- بن بست زمانی رخ می دهد که صفحات هیچ رابطه خروجی نداشته باشند.
- تغییر ضریب میرایی می تواند به همه ملاحظات بالا کمک کند. می توان آن را به عنوان احتمال اینکه یک وب گرد گاهی اوقات به یک صفحه تصادفی پریده و بنابراین در سینک گیر نمی کند تعبیر شود.

1. Old_Ax
2. LIGIRAN
3. Betflood
4. Mitingg
5. Tanzgaradagh
6. FOOTKHABAR_85
7. Betcart
8. Betforward
9. football_turk
10. Mind_C0ld

رتبه بندی صفحات بدین معناست که اگر کاربر از یک صفحه یا کانال شروع به مرور کند و بصورت تصادفی به لینک‌های موجود در صفحه برود چقدر احتمال دارن به یک صفحه یا کانال خاص برسد. بصورت کلی اهمیت صفحات(کانال) را مشخص میکند.

سوال سوم

پلیس فتا میخواهد کانالهایی را که در موضوع شرط بندی باعث کلاهبرداری و خروج پول از کشور میشوند را شناسایی کند. حال نکته اینجاست که تعداد چنین کانالهایی کم نیست و بررسی همه این کانالها و کنترل آنها به وقت زیادی احتیاج دارد. معیار مرکزیتی پیشنهاد دهید که با ترکیب شاخصها و اطلاعات مختلف بتواند کانالها را به ترتیبی متناسب با احتمال تاثیرگذاری بر روی کاربران مرتب کند.

$$\frac{\frac{views}{count}}{\sum_{i=[1,..,n]} \frac{views_i}{counts_i}} * PR * Auth$$

فرمول پیشنهادی برای رتبه بندی کانال‌ها را به صورت بالا تعریف میکنیم. این فرمول در قسمت اول متشکل از نرمال شده تعداد بازدید تقسیم بر تعداد پست مربوط به شرط‌بندی کانال میباشد. این قسمت ابتدا با تقسیم تعداد بازدید بر تعداد مطلب تخمین نرمال احتمالی بازدید هر مطلب منتشر شده در کانال را محاسبه میکند و سپس این مقدار بر اساس تمامی کانالها نرمال میشود تا تاثیر بازدید fake کم اثر بشود. سپس این مقدار در page rank کانال ضرب میشود تا احتمال اینکه کاربر با شروع از یک کانال و جستجو در میان مطالب فوروارد شده و رفتن به لینک‌های تصادفی به این کانال برخورد کند و همچنین اهمیت کانال در میان کانال‌های دیگر محاسبه شود. در مرحله آخر نیز اعتبار کانال توسط مرکزیت authority مورد بررسی قرار داده میشود تا بدانیم آیا کانال یک مرجع برای کانال‌های دیگر بوده یا خیر و مرجعیت آن نیز مورد بررسی قرار داده شود. زیرا با توجه به تعاریفی که برای یالها در این گراف داشتیم گره‌های هاب کانال‌هایی هستند که بیشتر مطالب آنها کپی از دیگر کانال‌ها هستند و مرجعیت و اعتبار پایینی دارند.

پایان.