تعریف و تشخیص اجتماعات در شبکه‌ها

فیلیپو ردیسیچی، کلودیو کاستلانو، فدریکو سیسکنی، ویتوریو لورتو، دومنیکو پاریسی

تشخیص ساختارهای اجتماعاتی یک موضوع مهم در بسیاری از زمینه‌ها می‌باشد. این موضوع با مفاهیمی مانند شبکه‌های اجتماعی(روابط بین اعضا)، تحقیقات بیولوژیکی یا مسایل تکنولوژیکی(بیهنه‌سازی زیرساخت‌های حجیم) در ارتباط می‌باشد.الگوریتم‌های متنوعی برای یافتن ساختارهای انجمنی موجود می‌باشد ولی، یک تعریف واحد و عمومی از اجتماعات در این الگوریتم‌ها ارایه نشده است که سبب می‌شود بدون داشتن اطلاعات غیرتپولوژیکالی، بسختی میتوان جواب‌ها را تشخیص داد. در این مقاله ما سعی می‌کنیم که تعاریف عمومی و کمی اجتماعات را که در الگوریتم‌ها استفاده شده‌اند را بیان نماییم. با این کار الگوریتم‌های تشخیص اجتماعات کاملا درک می‌شوند. در ادامه، ما یک الگوریتم محلی برای یافتن اجتماعات معرفی می‌کنیم که با در نظر گرفتن پیچیدگی محاسباتی، همان سطح دقت را حفظ می‌نماید. الگوریتم برروی شبکه‌های مصنوعی و شبکه‌های واقعی تست شده‌است. در حالت خاص نیز، الگوریتم را بروی شبکه‌ی همکاری دانشمندان نیز اجرا می‌نماییم، این شبکه به دلیل ابعادش کمتر مورد بررسی قرار گرفته است. این دسته از الگوریتم‌های محلی می‌توانند مسیر برای تحلیل سیستم‌های فوق حجیم تکنولوژیکی و بیولوژیکی باز نمایند.

در سال‌های اخیر شواهد با سرعت زیادی رشد نموده‌اند که بسیاری از سیستم‌های موجود در زمینه‌های مختلف را میتوان با شبکه‌های مدل نمود؛ بعنوان مثال به وسیله‌ی چند راس و یال و با خواص ساختاری کلی. سیستم‌های تکنولوژیکی مانند اینترنت یا بیولوژیکی مانند شبکه‌های متابیولوکی و سیستم‌های اجتماعی نمونه‌ای از این سیستم‌ها می‌باشند.

در این مقاله ما به ویژگی‌های ساختاری شبکه‌ها توجه می‌کنیم، ساختارهای انجمنی، که اخیرا بسیار مورد توجه قرار گرفته‌اند. مفهوم انجمن مشخص است و به مفاهیمی مانند استخراج اطلاعات یا دخیره‌سازی مرتبط می‌شود. از این زاویه تعریف انجمن عمومی می‌شود و با توجه به زمینه‌ی استفاده شده، با مفاهیمی مانند ماژولاریتی، کلاس، گروه، خوشه و غیره مشابه است. در میان زمینه‌های مختلفی که به این مفهوم مرتبط هستند، از همه با ارزشمندتر، ماژولاریتی شبکه‌های متابیولوکی و تشخیص انجمن‌ها در تارنامه‌های می‌باشد. این موضوع آخر با مفاهیمی مانند موتور جست‌وجوی نسل جدید، فیلترینگ متن[1] و دسته‌بندی خودکار[2] مرتبط است.

با اشاره به مسایل مرتبط، ارایه‌ی یک الگوریتم برای تشخیص انجمن‌ها در شبکه‌های عمومی بسیار حیاتی است. این عمل، به هر حال بسیار کلی می‌باشد.

از لحاظ کیفی، اجتماع زیرمجموعه‌ای از رئوس یک گراف است که ارتباطات داخلی بین رئوس حجیمتر از ارتباط با بقیه شبکه می‌باشد. یافتن اجتماعات در گراف را میتوان بطور عمومی به نگاشت شبکه به یک درخت تعبیر نمود (شکل ۱). در این درخت برگ‌ها راس‌ها می‌باشند که بوسیله‌ی شاخه‌ها، راس‌ها یا گروه‌های از رئوس بهم متصل شده‌اند؛ این یک ساختار تودرتو از اجتماعات می‌باشد.

الگوریتم‌های متنوعی برای یافتن این اجتماعات در مقالات معرفی شده‌اند. روش سنتی به نام خوشبندی سلسله مراتبی وجود دارد. برای هر جفت i وj راس در شبکه، وزن بین i و j محاسبه شده، که نشان دهنده‌ی نزدیکی این دو راس می‌باشد. در حالت شروع هیچ یالی وجود ندارد و فقط مجموعه‌ای از راس‌های منفرد وجود دارند. با این روش راس‌ها به مرور گروه‌های بزرگتر و بزرگتری که اجتماعات هستند را تشکیل می‌دهند، در نهایت درخت تبدیل به ریشه می‌شود که کل شبکه می‌باشد. این دسته از الگوریتم‌ها را تجمیع می‌نامند.

در دسته‌ای دیگر از الگوریتم‌ها که تقسیمی نام دارند، ترتیب تولید درخت برعکس می‌باشد: با تمام گراف شروع می‌شود و یال‌ها را حذف می‌کند؛ این روش سبب می‌شود که گروه‌های کوچکتر و کوچکتری که

---

[1] Content filtering

[2] Automatic classification

اجتماعات هستند تولید شوند. نکته مهم در این روش‌ها، تشخیص یالی است که بین اجتماعات قرار دارد و داخل اجتماع نباشد. اخیرا گیروان و نیومن[۳] یک الگوریتم تقسیمی ارایه داده‌اند که انتخاب یال‌ها برمبنایی مرکزیت بینابینی[۴] است. کلیت مرکزیت بینابینی توسط آنتونیس[۵] و فریمن[۶] ارایه شده است. با در نظر گرفتن تمام کوتاه‌ترین مسیرها بین تمام زوج راس‌ها، بینابینی یک یال برابر است با تعداد مسیرهای که از این یال عبور کرده‌اند. واضح است که زمانی که شبکه از تعدادی خوشه تشکیل شده باشد که توسط چند یال این خوشه‌ها بهم متصل هستند، تمام کوتاه‌ترین مسیرهای بین خوشه‌ها از این یال‌های واسط عبور خواهند کرد و سبب می‌شود امتیاز این یال‌ها بالا باشد. هر قدم این الگوریتم یافتن امتیاز بینابینی برای تمام یال‌ها و حذف بزرگترین‌ها می‌باشد. با اجرای این الگوریتم شبکه به زیرگراف‌های کوچکتر تقسیم می‌شود که خود این زیرگراف‌ها نیز در ادامه تقسیم خواهند شد تا در نهایت شبکه به راس‌های منفرد تقسیم شود. در این روش درخت از ریشه به برگ تولید می‌شود.

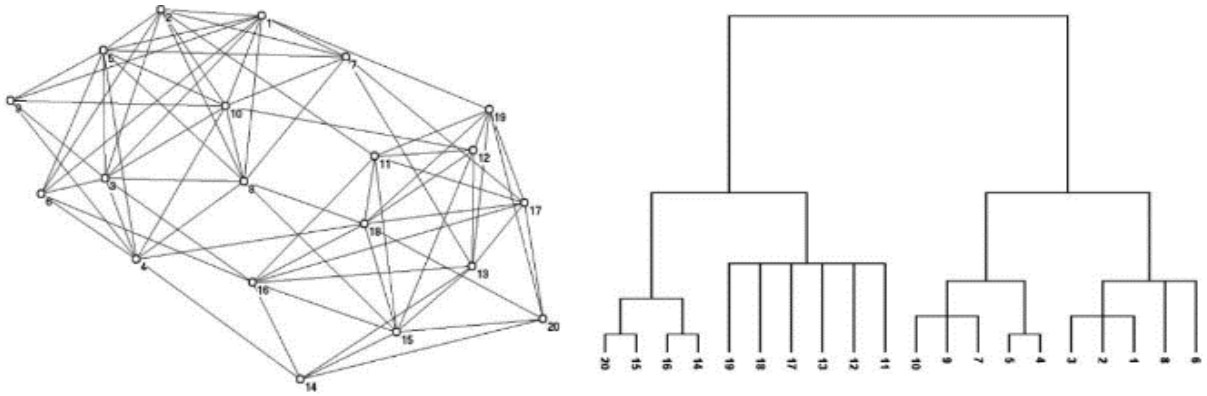
این الگوریتم یک قدم بزرگ در تشخیص اجتماعات برداشته است زیرا، بسیار از قصورات روش‌های سنتی را ندارد. این حقیقت، بیانگر این است که چرا این الگوریتم در چند سال اخیر بعنوان یک روش استاندارد در تحلیل انجمن‌ها مورد استفاده قرار گرفته است.

این مقاله مسیرهای مختلفی را دنبال می‌کند تا یک روش جایگزین برای یافتن اجتماعات ارایه دهد. این روش مکمل برای پاسخ به دو نیاز زیر ارایه می‌شود.

۱. در کل، الگوریتم‌ها اجتماعات را با توجه به اینکه چه چیزی پیدا می‌کنند تعریف می‌نمایند. در درخت، بعنوان مثال، اجتماعات همیشه توسط الگوریتم‌ها بدون توجه به تحلیلی بروی ساختار شبکه تشخیص داده می‌شوند. این قصور ناشی از عدم وجود یک روش مشخص برای تشخیص اجتماع‌پذیر یا اجتماع‌ناپذیر بودن شبکه است. در نتیجه در شبکه‌های خاص، نیاز به دانش غیر تپولوژیکالی از ماهیت شبکه داریم تا متوجه شد که کدام شاخه از درخت کاراتر است. بدون این دانش، تشخیص اینکه انجمن صحیح است واضح نخواهد بود. دو مقاله‌ی قابل ملاحضه در این زمینه ارایه شده است. در روش پیشنهادی توسط ویکسون[۷] و هابرمن[۸] به کوچکترین سطح ساختار اجتماعی محدود شده است و برای

---

[۳] Girvan Newman
[۴] Edge Betweenness
[۵] Anthonisse
[۶] Freeman
[۷] Wilkinson
[۸] Huberman



شکل ۱ یک شبکه‌ی ساده و درخت مرتبط با این شبکه

الگوریتم‌های بینابینی تعریف شده است. اخیرا نیومن و گیروان، مقیاسی برای اندازگیری همبستگی اجتماعات بنام ماژولاریتی معرفی کرده‌اند. بشکل دقیقتر، ماژولاریتی نسبت لینک‌های داخلی در یک اجتماع را نسبت به حالتی که یال‌ها تصادفی بین رئوس رسم شوند را محاسبه می‌کند. این کمیت، یک مقیاس مناسب برای تشخیص همبستگی در اجتماعات را ارایه می‌دهد؛ بدلیل اینکه تعریف دقیق از کمیت اجتماع وجود ندارد، سبب شده است که نتوان تبعیضی بین اجتماعات معنادار قایل شد.

۲. الگوریتم بینابینی ارایه شده توسط نیومن و گیروان بسیار هزینه‌بر می‌باشد. محاسبه‌ی امتیاز تمام یال‌ها مرتبه زمانی برابر با ضرب تعداد یال‌ها (m) در تعداد راس‌ها (n) دارد. تکرار این محاسبات برای حذف تمام یال‌ها، سبب می‌شود که مرتبه‌ی بدترین حالت این الگوریتم برابر با m^2n باشد که سبب می‌شود این الگوریتم حتی برای شبکه‌های سایز متوسط (۱۰،۰۰۰ راس) نیز ناکارا باشد.

در این مقاله ما راه حلی برای هر دو مشکل ارایه می‌دهیم. در ابتدا یک ملاک برای اینکه کدام یک از زیر گراف‌های یافته شده توسط الگوریتم، یک انجمن می‌باشد ارایه می‌دهیم. دو تعریف کمی از مفهوم انمجن را بدقت بررسی خواهیم‌کرد. با اینکار، الگوریتم نیومن-گیروان را جامع‌تر خواهیم نمود. در قدم دوم، یک الگوریتم جایگزین برمبنای معیارهای محلی ارایه خواهیم داد که دقتی مشابه با الگوریتم نیومن-گیروان دارد ولی از بعد مرتبه اجرا بسیار کاراتر از این الگوریتم می‌باشد. شایان به ذکر است که بعد از اتمام این مقاله، نیومن بعنوان یک الگوریتم سریعتر به این الگوریتم ارجاع می‌نماید.

تعریف کمی انجمن

با یک ایده‌ی ساده میتوان اولین مشکلی که مطرح شده است را حل نمود: الگوریتمی که درخت را تولید می‌کند صرفا زیرگراف‌های که

مناسب انمجن‌پذیری هستند را انتخاب نماید. در قدم بعد صحت انجمن بودن یا نبودن این زیرگراف‌ها را بررسی می‌نماید. اگر این زیر گراف ملاک‌های انجمنیت را نداشته باشد پس شاخه‌ی مربوط به این زیر گراف در درخت نباید گسترش یافته و رسم شود.

همانگونه که قبلا اشاره شده است انجمن یک زیر گراف از شبکه می‌باشد که تعداد یال‌های داخلی بین راس‌های زیر گراف نسبت به یال‌های خارجی از زیرگراف بیشتر باشد. برای افزایش دقت الگوریتم‌ها لازم است که تعریف دقیقتری برای انجمن ارایه دهیم. تعاریف معقول و متنوعی در مقالات ارایه شده است. در این بخش با توجه با الگوریتم‌های موجود، یک تعریف محتمل از انجمن را به شکل فرمول بیان می‌کنیم.

k_i یک کمیتی است که درجه‌ی راس i را نشان می‌دهد که برابر با جمع سطر iام از ماتریس مجاورت گراف است. اگر یک زیرگراف بنام V را در نظر بگیریم که راس i عضو این زیر گراف باشد، میتوان نوشت:

$$k_i(V) = k_i^{in}(V) + k_i^{out}(V)$$

که $k_i^{in}(V)$ برابر با یال‌های از راس i است که داخل زیر گراف V قرار دارند و $k_i^{out}(V)$ برابر است با تعداد یال‌های از راس i که داخل زیر گراف V قرار ندارند.

حال زیرگراف V یک اجتماع قوی[۹] است اگر داشته باشیم:

$$k_i^{in}(V) > k_i^{out}(V).\ \ \forall i \in V \quad (۱)$$

زیرگراف V یک اجتماع ضعیف[۱۰] است اگر داشته باشیم:

$$\sum_{i \in V} k_i^{in}(V) > \sum_{i \in V} k_i^{out}(V) \quad (۲)$$

در اجتماع ضعیف جمع تعداد یال‌های داخل اجتماع از جمع تعداد یال‌های خارج شده از اجتماع بیشتر است.

هر اجتماع قوی لزوما یک اجتماع ضعیف نیز می‌باشد ولی عکس این موضوع صادق نیست.

شایان بذکر است که تعریف ما از اجتماع، طبیعتا تنها تعریف ممکن نمی‌باشد. تعاریف متعدد دیگری که حتی در بعضی از زمینه‌ها قویتر بیان شده‌اند نیز وجود داشته و چاپ شده‌اند. در میان این تعاریف، تعریف گردایه‌ی ال-اس[۱۱] همراستا با تعریف اجتماع قوی می‌باشد البته کمی سخگیرانه‌تر. گردایه‌ی ال-اس زیر مجموعه‌ای است که هر زیر مجموعه از آن با بقیه‌ی زیر مجموعه نسبت به مابقیه گراف همبستگی

---

[۹] Strong Community
[۱۰] Weak Community
[۱۱] LS-Set

بیشتری داشته باشد. از سوی مفهوم کی-هسته[۱۲] نیز تقریبا مشابه با تعریف اجتماع ضعیف می‌باشد. در زیرگراف کی-هسته هر راس حداقل به کا راس دیگر از زیرگراف متصل است.

الگوریتم جامع

با توجه به تعاریف بالا، اگر شبکه‌ای به دو بخش تقسیم شود بگونه‌ای که یک بخش بزرگ باشد و بخش دیگر شامل تعدادی کم راس باشد، بخش بزرگتر همواره ملاک اجتماع را ارضا می‌کند. برای تحلیل این مشکل، گراف مصنوعی اردوش-رنه[۱۳] را در نظر می‌گیریم. اگر شبکه را به گونه‌ای تقسیم نماییم که در یک قسمت αN راس و در قسمت دیگر $N(1-\alpha)$ راس قرار داشته باشند، بوسیله مفاهیم احتمالی می‌توان اثبات کرد که به احتمال $P(\alpha)$ بخشی که شامل αN راس است تشکیل اجتماع ضعیف یا قوی می‌دهد. زمانی که تعداد راس‌ها بسیار زیاد باشد مقدار احتمال به ۰٫۵ نزدیک می‌شود. با توجه به تعاریف فوق پس در یک گراف تصادفی، زمانی که شبکه را تصادفا به دو بخش تقسیم نمایند، قسمت بزرگتر تشکیل اجتماع می‌دهد. در هر حال بسیار بعید است که هر دو زیرگراف همزمان ملاک را رعایت نمایند؛ حال اگر تقسیم بشرطی که هردو قسمت ملاک را رعایت نمایند پذیرفته شود پس این نتیجه‌ی تایید شده حاصل می‌شود که شبکه‌ی تصادفی اجتماع‌پذیر نیست. این ملاک را به حالت کلی گسترش می‌دهیم: بعد از تقسیم اگر حداقل دو زیرگراف همزمان ملاک تعریف شده را رعایت ننمایند پس تقسیم نادرست می‌باشد.

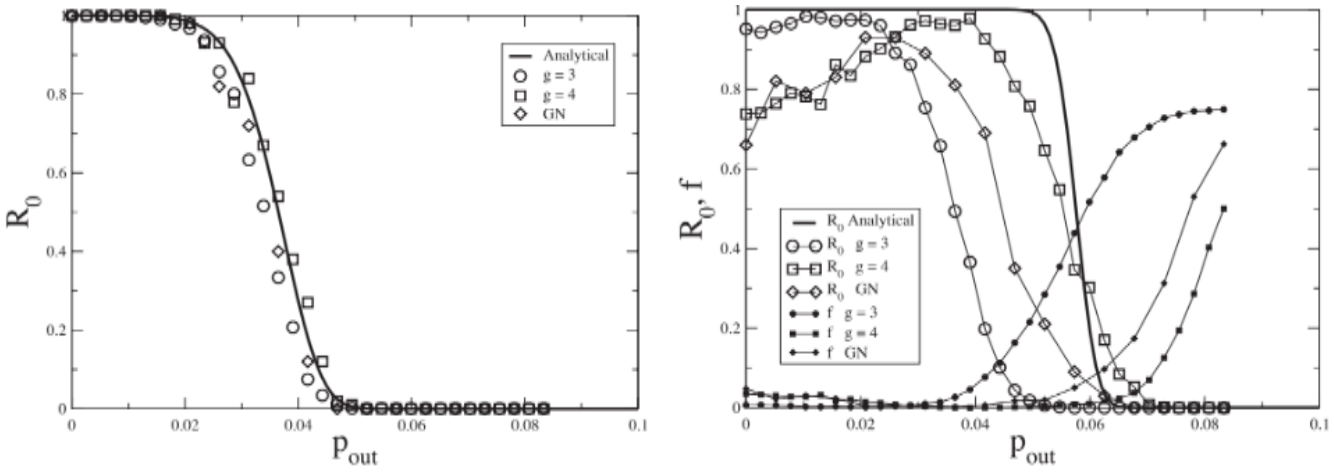
حال می‌توانیم الگوریتم بهبود یافته‌ی نیومن-گیروان را خلاصه بیان نماییم:

۱- تعریف اجتماع را انتخاب کن.
۲- مرکزیت بینابینی را برای تمام یال‌ها محاسبه کن و آنهای که بیشترین مقدار را داند حذف کن.
۳- اگر حذفیات سبب نشد که گراف تقسیم شود برو به قدم ۲.
۴- اگر حذفیات گراف را تقسیم کرده باشد، بررسی کن که حداقل دو زیرگراف ملاک تعریف شده را رعایت نمایند. اگر شد پس قسمت مربوط را در درخت رسم کن.
۵- رویه را تا زمانی (با رفتن به نقطه ۲) که تمام یال‌ها حذف نشده‌اند ادامه بده.

بسیار مهم است که یادآوری شد که کمیت‌های تعریف شده در معادلات ۱ و ۲ باید با توجه به ماتریس مجاورت ارزیابی شوند. نتیجه‌ی اجرای این الگوریتم بروی یک شبکه، درختی خواهد بود که هر شاخه‌ی آن با توجه به ملاک تعیین شده یک اجتماع معنی‌دار می‌باشد.

---

[۱۲] K-Core
[۱۳] Erdos-Renyi



شکل ۲ ارزیابی الگوریتم‌های مختلف در گراف مصنوعی که چهار اجتماع دارد. نحوی تولید شبکه در متن شرح داده شده است. تعداد راس‌ها برابر با ۱۲۸ است و مقدار p_{in} و p_{out} بگونه‌ای در تغییر هستند که متوسط درجه برابر با شانزده باشد. سمت چپ اجتماعا قوی هستند که نسبت دقت الگوریتم‌های مختلف را با مقادیر بدست آمده از تحلیل احتمالی برای چهار انجمن مقایسه می‌کند. شکل سمت راست نیز همین مقایسه را برای اجتماعات ضعیف نشان می‌دهد. برای هر الگوریتمی تعدادی از راس‌ها بدرستی دستبندی نشده‌اند.

حال می‌توان کارایی الگوریتم نیومن-گیروان را ارزیابی نمود. ما برای این کار از گراف مصنوعی که توسط نیومن و گیروان ارایه شده است استفاده می‌کنیم. یک گراف ساده به اندازه N که به ۴ بخش تقسیم شده است: ارتباط بین زوج‌های داخل یک اجتماع با احتمال p_{in} نشان داده می‌شوند، در حالی که احتمال بین زوجی که راس‌های آن در دو گروه مختلف است را با احتمال p_{out} نشان داده شده است. هر قدر که احتمال p_{out} بیشتر شود اجتماعات ضعیفتر می‌شوند.

برای هر گراف مصنوعی، خروجی الگوریتم یک درخت می‌باشد. اگر الگوریتم بتواند چهار عدد اجتماع را پیدا نماید، هر راس در اجتماع درست قرار بگیرد و زیرگراف‌ها تقسیم‌پذیر نباشند پس موفق عمل کرده است.

شکل (۲) نسبت موفقیت الگوریتم نیومن-گیروان را به مقادیر مورد نظر که بصورت تحلیلی بدست آمده‌اند، مقایسه می‌کند. همانگونه که مشخص است الگوریتم اجتماعات قوی را بخوبی تشخیص می‌دهد ولی اجتماعات ضعیف را بخوبی تشخیص نمی‌دهد. به هرحال کمیت‌های شکل (۲) نباید باعث گمراهی شوند. با ساده در نظر گرفتن ملاک موفقیت، الگوریتم در اجتماعات ضعیف چهار اجتماع یافته است ولی تعدادی محدود از راس‌ها را به اشتباه در اجتماعات دیگر قرار داده است. این راس‌ها دارای احتمال p_{out} بالایی می‌باشند. انحرافاتی در حالت تیوریکال مشاهده می‌گردد که ناشی از این است که مقدار p_{out} بسیار کم باشد؛ به این دلیل که خود این چهار اجتماع به اجتماعات کوچکتری تقسیم می‌شوند. این حادثه در محاسبات تحلیل مدنظر گرفته نشده است اما اگر

اندازه‌ی سیستم بزرگ شود قابل ملاحضه می‌شوند.

یک الگوریتم سریع

الگوریتم نیومن-گیروان بسیار هزینه‌بر می‌باشد زیرا ارزیابی‌های متعدد برای هر یال در شبکه انجام می‌شود که نیازمند متریک‌های جهانی هستند، مرکزیت بینابینی، که برپایه‌ی ویژگی‌های کل سیستم می‌باشد. با اینکه روش‌های زیرکانه‌ای برای محاسبه‌ی مرکزیت بنیابینی بصورت همزمان ارایه شده است اما هنوز هم بخش پرهزینه‌ای در الگوریتم می‌باشد. در نتیجه، زمان مورد نیاز برای اجرای کامل الگوریتم با توجه به اندازه‌ی مسله، بسیار سریع رشد می‌کند و سبب می‌شود این الگوریتم برای گراف‌های بالا ۱۰،۰۰۰ راس ناکارا باشد.

برای برطرف نمودن این مشکل، ما یک الگوریتم تقسیم کننده که برپایه‌ی کمیت‌های محلی می‌باشد معرفی می‌کنیم که از الگوریتم نیومن-گیروان سریعتر می‌باشد. بخش اصلی الگوریتم‌های تقیسم کننده کمیتی است که با آن بتوان یال‌های بین اجتماعات را تشخیص داد. ما یک مفهوم بنام ضریب خوشه‌ای یال[۱۴] را معرفی می‌کنیم که مشابه ضریب خوشه‌ای راس[۱۵] می‌باشد. این کمیت برابر است با تعداد مثلث بسته‌های که یال مدنظر در آنها وجود دارد بر تعداد مثلث بسته‌ای که یال مدنظر بالقوه می‌تواند تشکیل دهد که برابر با درجه راس مربوط به یال می‌باشد. به بیان رسمی‌تر مقدار ضریب خوشه‌ای یالی که از راس i به راس j رسم شده است برابر است با:

$$C_{i.j}^{(۳)} = \frac{z_{i.j}^{(۳)}}{\min[(k_i - ۱).(k_j - ۱)]} \qquad (۳)$$

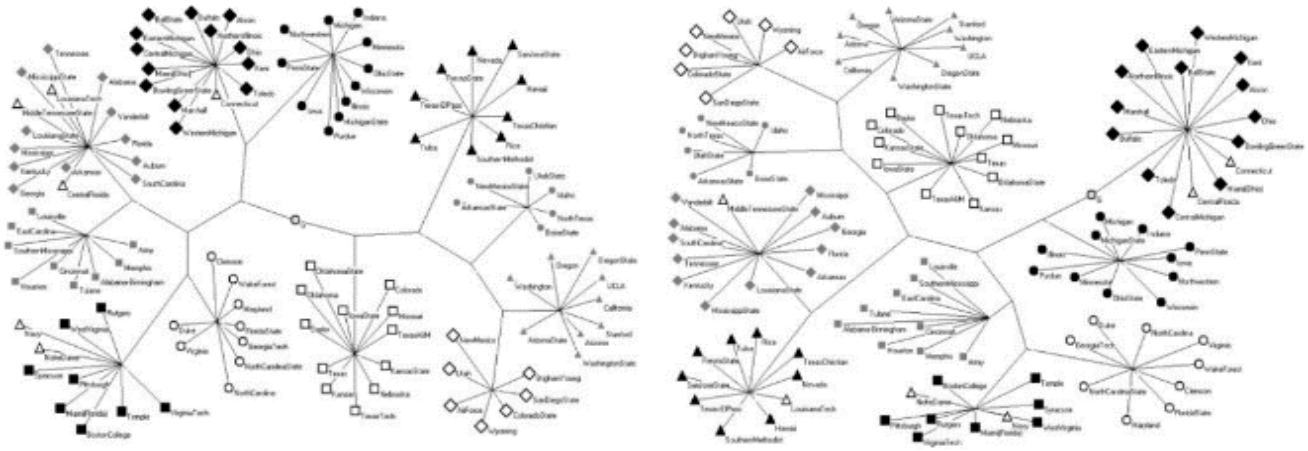
که $z_{i.j}^{(۳)}$ برابر با تعداد مثلث‌های موجود بروی یال می‌باشد و مخرج کسر نیز حداکثر مثلث ممکن بروی این یال می‌باشد.

ایده پشت این متریک برای استفاده در الگوریتم تقسیم کننده این است که یال‌های که دو اجتماع را بهم متصل می‌نماید تعداد کمی مثلث یا هیچ مثلثی تشکیل نمی‌دهند و مقدار $C_{i.j}^{(۳)}$ برای آنها کم می‌باشد. از طرفی یال‌های داخل خوشه نیز مقدار $C_{i.j}^{(۳)}$ بالایی دارند. پس کمیت $C_{i.j}^{(۳)}$ یک معیاری است که میزان داخل اجتماع بودن یک یال را نشان می‌دهد. مشکل زمانی ایجاد می‌شود که تعداد مثلث‌ها صفر باشد که سبب می‌شود مقدار $C_{i.j}^{(۳)}$ جدا از مقدار مخرج همواره صفر باشد. این مشکل با افزودن یک مقدار ثابت به صورت کسر برطرف می‌شود پس داریم:

---

[۱۴] Edge Clustering Coefficient

[۱۵] Vertex Clustering Coefficient



شکل ۳ درخت رسم شده برای گراف تیم فوتبال دانشگاهی (چپ) الگوریتم نیومن-گیروان و (راست) الگوریتم ما با مقدار $g=4$ علایم مختلف بیانگر تیم‌های هر کنفرانس مختلف است. در هردو شکل، اجتماعات بدرستی متناسب با کنفرانس‌ها، تشخیص داده شده‌اند. البته شش تیم به اشتباه دستبندی شده‌اند.

$$C_{i.j}^{(۳)} = \frac{z_{i.j}^{(۳)} + ۱}{\min[(k_i - ۱).(k_j - ۱)]} \qquad (۴)$$

همچنین می‌توان این معادله را برای دورهای با طول بزرگتر (g) نیز نوشت:

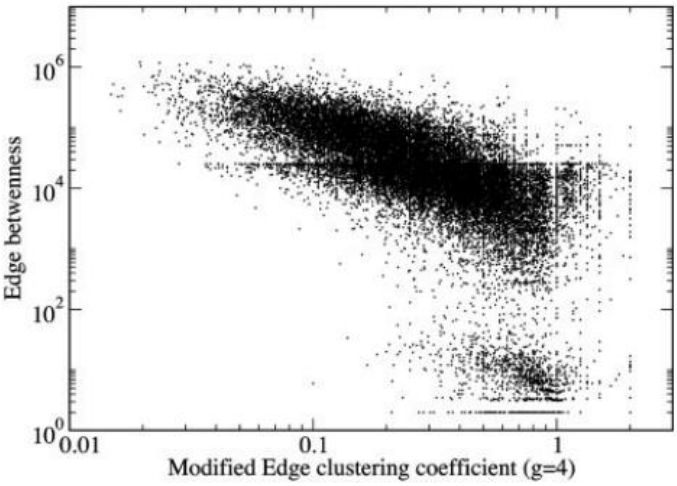
$$C_{i.j}^{(g)} = \frac{z_{i.j}^{(g)}}{s_{i.j}^{(g)}} \qquad (۵)$$

که $z_{i.j}^{(g)}$ برابر با تعداد دورهای بطول g است که راس (i,j) در آن قرار دارد و $s_{i.j}^{(g)}$ برابر با تعداد دورهای بالقوه‌ای طول g است که با این یال می‌توان ساخت.

حال می‌توان با توجه به مقادیر مختلف g یک الگوریتم تشخیصی مشابه با الگوریتم نیومن-گیروان ارایه کرد که در هر دور از اجرا، یال‌های که کمترین مقدار $C_{i.j}^{(g)}$ دارند را حذف می‌کند. با توجه با مقدار g می‌توان الگوریتم را بصورت محلی یا غیر محلی اجرا نمود. لازم بذکر است که با توجه به تعریف $C_{i.j}^{(g)}$ راسی که فقط یک یال دارد، تضمین می‌شود که بعنوان یک انجمن تنها درنظر گرفته نمی‌شود زیرا مقدار $C_{i.j}^{(g)}$ برای یال منفرد برابر با بینهایت می‌شود.

دقت این الگوریتم را با مقایسه با الگوریتم نیومن-گیروان بررسی کرده‌ایم که در شکل (۲) نشان داده شده است. نتیجه‌ی حاصل شده این است که این الگوریتم در مقادیر g=٤ و g=٣ برای انجمن‌های قوی به اندازه‌ی نیومن-گیروان دقیق می‌باشد.

از طرف دیگر، برای انجمن‌های ضعیف نیز با مقدار g=٤ بهترین نتیجه حاصل می‌شود. ارزیابی دیگری نیز بروی شبکه‌ی اجتماعی که توسط الگوریتم نیومن-گیروان مورد بررسی قرار گرفته است، صورت گرفت. شکل (۳) درخت حاصل شده برای گراف تیم فوتبال دانشگاهی را که توسط الگوریتم نیومن- گیروان و الگوریتم ما با مقدار g=٤ تولید شده است را نشان می‌دهد.



شکل ۴ مرکزیت بنابینی در مقابل ضریب خوشه‌ای یال برای شبکه‌ی همکاری دانشمندان در ساختار انجمنی شبکه‌ی همکاری داشنمندان. هر نقطه نشانگر یک یال در شبکه است. برای جزییات بیشتر ساختار انجمنی در شبکه‌ی همکاری دانشمندان را مشاهده نمایید

نتایج بسیار مشابه بهم هستند که بیانگر این است که الگوریتم محلی انجمن‌ها را بخوبی تشخیص می‌دهد.

مقایسه عمیقتری بین الگوریتم نیومن-گیروان و ضریب خوشه‌ای یال در شکل (۴) نشان داده شده است. مقایسه بین مرکزیت بینابینی و مقدار $C_{i.j}^{(۴)}$ برای هر یال از گراف همکاری بین دانشمندان می‌باشد. واضح است که انطباقی بین این دو متریک وجود دارد؛ یال‌های با مرکزیت بینابینی بالا، مقدار $C_{i.j}^{(۴)}$ کمی دارند. تطابق کامل نیست؛ یالی که کمترین مقدار $C_{i.j}^{(۴)}$ را دارد، بیشترین مقدار مرکزیت بینابینی را ندارد. در نهایت ما انتظار داریم که هر دو الگوریتم نتایج مشابه‌ای ارایه دهند که البته لزوما کاملا منطبق نباشند.

در این چهارچوپ لازم است که به مقایسه‌ی بارت[۱۶] که بین مرکزیت بینابینی و افزونگی[۱۷] می‌باشد اشاره نماییم. مفهوم افزونگی بسیار مشابه با خوشبندی راس می‌باشد و مشابه با روح کار ما است. بارت اشاره دارد که راس‌های که به چند دور تعلق دارند مرکزیت بینابینی بالایی دارند.

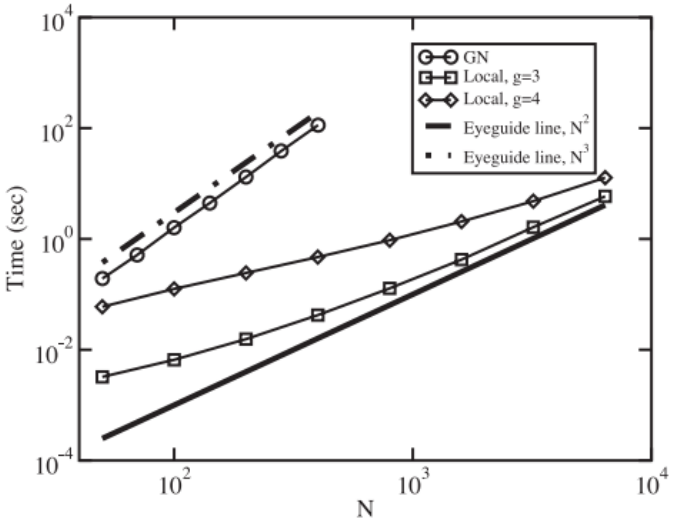
اکنون می‌توانیم کارایی محاسباتی الگوریتم محلی را شرح دهیم. می‌توان پیچیدگی محاسبات را بشرح زیر تقریبا تخمین زد. زمانی که یک یال حدف می‌شود باید بررسی شود که آیا شبکه قسمت شده است و همچنین مقادیر $C_{i.j}^{(g)}$ نیز در همسایگی یال حذف شده باید بروز شوند. بخش اول پیچیدگی برابر با تعداد تمام یال‌ها دارد اما بخش دوم در مرتبه‌ی m انجام‌پذیر نیست. زیرا این عمل برای تمام یال‌ها باید تکرار شود، پس میتوان انتظار داشت که مرتبه زمانی برابر با $am + bm^۲$ باشد. ما انتظار داریم برای شبکه‌های کوچک پیچیدگی برابر با m باشد و برای شبکه‌های بزرگ مرتبه برابر با $m^۲$ است.

سرعت این الگوریتم را با اجرا بروی شبکه‌ای تصادفی که نرخ رشد راس آن N و میانگین درجه راس آن ثابت است، بررسی نمودیم. نتایج در شکل (۵) نشان داده شده‌اند که بیانگر این است که الگوریتم برپایه‌ی ضریب خوشه‌ای، هم

---

[۱۶] Burt

[۱۷] redundancy



شکل ۵ زمان مورد نیاز برای تحلیل شبکه تصادفی به اندازه‌ی N که متوسط درجه‌ی آن ۵ است. زمان مورد نیاز برای تقسیم شبکه از کل به راس‌های منفرد محاسبه شده است. هیچ ملاکی برای اتمام در نظر گرفته نشده است. سیستم استفاده شده پردازنده‌ی ۸۰۰ مگاهرتزی دارد.

برای g=۳ و g=٤، سریعتر از الگوریتم برپایه‌ی مرکزیت بینابینی است. کاملا مشهود است که سرعت رشد محاسبات برای g=۳ متناسب با N و در نهایت متناسب با N^2 است.

تشخیص اجتماعات در شبکه‌ی همکاری دانشمندان

در این بخش عملکرد الگوریتم سریع را بروی شبکه‌ی همکاری دانشمندان بررسی می‌کنیم. شبکه‌ی مدنظر ما دانشمندانی هستند که در بازه‌ی زمانی ۱۹۹۹-۱۹۹۵ مقاله‌ای در آرشیو یی-پرینت[۱۸] مربوط به کاندس مدر[۱۹] ثبت کرده باشند[۲۰]. این داده توسط مارک نیومن در اختیار ما قرار گرفت.

این شبکه از ۱۵،۶۱۶ راس(دانشمند) تشکیل شده است که یک مولفه‌ی بزرگ به اندازه‌ی ۱۲،۷۲۲ راس دارد. توجه ما به این مولفه‌ی بزرگ است و آن را بعنوان ورودی به الگوریتم سریع با مقادیر g=٤ و g=۳ داده‌ایم. زمان مورد نیاز برای تولید درخت با مقدار g=۳ با یک کامپیوتر شخصی که پردازنده‌ی آن ۸۰۰ مگاهرتزی است، برابر با ۳ دقیقه می‌باشد. الگوریتم در این زمان هم اجتماعات ضعیف و هم اجتماعات قوی را تشخیص داده است. در پایان رویه، برای هر مقدار g لیستی از اجتماعات ضعیف و لیستی از اجتماعات قوی تولید می‌گردد. شکل (۶) توزیع اندازه‌ی اجتماعات ضعیف را نشان می‌دهد. این نمودار نشان دهنده‌ی تبعیت اندازه‌ی اجتماعات از قانون توان است که معادله آن $P(S) \approx S^{-t}$ $t \approx ۲$ می‌باشد. توانی که ما یافتیم با توانی که توسط الگوریتم نیومن-گیروان محاسبه می‌شود یکی می‌باشد.

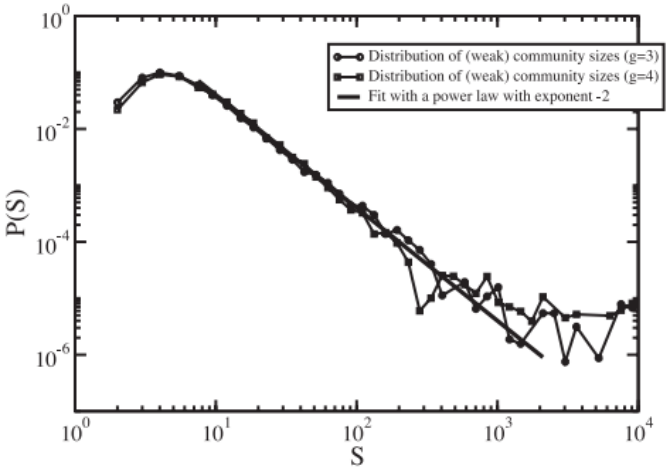
مسله مهم در تحلیل اجتماعات، نبود ابزاری است که بتوان با جزییات بیشتر انجمن‌ها را تحلیل نمود زیرا چنین کمیتی تعریف نشده است. بسیاری مستقیما از درخت استفاده

---

[۱۸] E-print Archive

[۱۹] Condensed Matter

[۲۰] http://xxx.lanl.gov/archive/cond-mat



شکل ۶ نرمالسازی توزیع اندازه‌ی تمام اجتماعات ضعیفی که توسط الگوریتم سریع با مقادیر $g=3$ و $g=4$ کشف شده است. در هر دو حالت رفتاری نزدیک به قانون توان با مقدار ۲- دراند

می‌کنند: آیا اجتماعات واقعا همکاری بین دانشمندان را بدرستی نمایش میدهند؟ آیا زمینه‌های تحقیقاتی را نشان می‌دهند؟ آیا دانشمندان خودشان قبول دارند که به این زمینه‌ی تحقیقاتی تعلق دارند؟ تمامی این سوالات به کمک کمیت‌های کمی و کیفی قابل پاسخ نیستند. ما این مسیر را پیموده‌ایم و زیرگراف‌های مختلف را در سلسله‌های متفاوت بررسی نموده‌ایم. در بهترین تحقیقات ما، نتایج قابل قبول بودند. اما این ادعا نباید این منظور را برساند که الگوریتم ما بهترین الگوریتم می‌باشد. ما ترجیح می‌دهیم که خود خواننده نتایج کار ما را ارزیابی نماید و ما نیز اطلاعات اضافی را درصورت درخواست در اختیار قرار می‌دهیم.

جمعبندی

تشخیص اجتماعات در شبکه‌های پیچیده‌ی بزرگ مبحثی است بسیار جای تحقیق دارد. مفهموم انجمن ملموس می‌باشد اما، برای تشخیص انجمن در شبکه‌ها باید متریک‌های کمی و نامبهم از مفهموم انجمن تعریف نماییم. زمانی که تعریفی درست ارایه شود می‌توان تمام زیرگراف‌های شبکه را بررسی نمود که تعریف انجمن را رعایت می‌کنند یا نه. این روش در عمل حتی برای شبکه‌های کوچک نیز غیر ممکن می‌باشد. به همین دلیل یافتن اجتماعات، هدفی محدود باید باشد: انتخاب زیرگراف‌های که درخت تعریف شده در بخش قبل را تولید می‌کنند. الگوریتم‌های تقیسم کننده و تجمیع کننده این عمل را انجام می‌دهند. مقایسه بین این دسته از الگوریتم‌ها بسیار کلی می‌باشد. در بعضی از مواقع به کمک گراف مصنوعی چهار بخشی که در قسمت قبل معرفی شده است می‌توان دقت جواب‌ها را مقایسه کرد. در مواقعی دیگر مانند زمانی که انجمن‌های شبکه‌ی دانشمندان را تشخیص می‌دهیم، متریکی وجود ندارد که دقت انجمن‌ها و درخت را محک بزند. معمولا فقط بررسی می‌شود که جواب قابل قبول باشد. به هر حال این قرارداد بسیار دور از هدف است و خود فرد ناظر دیدگاه خود از انجمن را در این تعریف دخیل می‌نماید.

در این مقاله ما دو روش مختلف برای تولید درخت را توضیح دادیم. در ابتدا ما معیاری برای تعریف انجمن ارایه دادیم که متناسب با الگوریتم‌های تقسیم کننده است. با این کار الگوریتم جامع می‌شود و بدون نیاز به اطلاعات

غیرساختاری قابل اجرا است. در قدم بعد یک الگوریتم تقسیم کننده و محلی ارایه شد که بسیار سریع است. هر دو این دست آوردها ارزیابی شده و نتایج مثبت داشتند. نتایج بدست آمده از تحلیل شبکه‌ی همکاری دانشمندان رضایت بخش است. البته با توجه به مباحث بالا این موضوع هنوز بسیار جای کار دارد و فعلا از این دقیقتر نمی‌توان عمل نمود. قویا ارایه یک متریک برای ارزیابی درخت از اهداف مهم در این بخش می‌باشد.

یک نکته قابل ذکر است که تا با الان ما فقط شبکه‌های اجتماعی را تحلیل نموده‌ایم. ساختار شبکه‌های اجتماعی با ساختار سایر شبکه‌ها مانند شبکه‌های بیولوژیکی و تکنولوژیکی، متفاوت است. در بین این تفاوت‌ها، عدم تطابق بین درجه راس در شبکه‌های غیر اجتماعی بسیار مشهود می‌باشد. با اینکه نتایج ما و سایر کارها در شبکه‌های اجتماعی می‌باشد، سوال اصلی این است که آیا در شبکه‌های غیر اجتماعی نیز الگوریتم‌ها کارا هستند یا نه. الگوریتم ما بر اساس دورهای کوچک بنا شده است و ممکن است در شبکه‌های غیر اجتماعی کارا نباشد زیرا تعداد دور کوچک در این شبکه‌ها کم می‌باشد. در یک آزمایش که بروی چهار شبکه صورت گرفت (دو شبکه‌ی اجتماعی و دو شبکه‌ی غیراجتماعی) متریکی بنام میانگین ضریب توری[۲۱] را محاسبه نمودند که مقدار این متریک دو تا چهار برابر مقدار همین متریک در نمونه شبکه‌ی تصادفی است که هم سایز با این شبکه‌ها می‌باشد. از این آزمایش می‌توان نتیجه گرفت که الگوریتم ما در شبکه‌های غیراجتماعی نیز کارا خواهد بود البته باید در این زمینه نیز تحقیقاتی صورت گیرد.

ما باور داریم که مفاهیم تعریف شده در این مقاله در تحلیل شبکه‌ها بسیار کمک کننده است. از طرفی ارایه متریکی که انجمن را قابل تعریف نماید سبب شده است که الگوریتم‌ها جامع باشند و به کمک اطلاعات ساختاری شبکه، اجتماعات را تشخیص دهند. در سمت دیگر نیز ارایه الگوریتمی محلی، مسیر را برای تحلیل سیستم‌های حجیم هموار می‌نماید.

ما از مارک نیومن بدلیل اینکه شبکه‌ی همکاری دانشمندان و گراف تیم فوتبال دانشگاهی را مهیا نمودند تشکر می‌نماییم. همچنین از آلین بارات[۲۲] بدلیل راهنمایی‌ها و نظراتشان سپاس گذاریم.

---

[۲۱] Average Grid Coefficient

[۲۲] Alain Barrat