فصل دهم

موضوع : Mining Social-Network Graphs

به نام خدا

مقدار زیادی داده با تجزیه و تحلیل از شبکه های اجتماعی به دست می آید.

شناخته شده ترین نمونه یک شبکه اجتماعی ، رابطه "دوستان" است که در سایت هایی مانند فیس بوک یافت می شود.

با این حال ، همانطور که خواهیم دید بسیاری از منابع داده دیگر وجود دارند که افراد یا نهادهای دیگر را به هم وصل می کنند.

در این فصل تکنیک های تحلیل داده های موجود در شبکه ها را بررسی می کنیم.

یک سوال مهم در مورد یک شبکه اجتماعی ، چگونگی شناسایی "اجتماعات" ، یعنی زیر مجموعه گره ها (افراد یا اشخاص دیگری که شبکه را تشکیل می دهند) با اتصالات غیرمعمول قوی است.

برخی از تکنیک های مورد استفاده برای شناسایی جوامع مشابه الگوریتم های خوشه بندی است که در فصل 7 درباره آنها گفتیم.

با این حال ، مجموعه ها تقریباً هرگز قسمت هایی از گره ها را در یک شبكه تقسیم نمی كنند.

در عوض ، مجموعه ها معمولاً با هم همپوشانی دارند.

به عنوان مثال ، شما ممکن است به چندین گروه مختلف از دوستان یا همکلاسی های خود تعلق داشته باشید.

مردم یک جامعه تمایل دارند یکدیگر را بشناسند ، اما مردم دو جامعه مختلف به ندرت یکدیگر را می شناسند.

شما نمی خواهید فقط در یکی از گروه ها قرار بگیرید و همچنین معقول نیست که همه ی افراد از همه ی جوامع خود را حتما در یک گروه قرار دهید.

همچنین در این فصل الگوریتم های کارآمد را برای کشف سایر خصوصیات گراف ها بررسی می کنیم.

ما به یک مفهومی تحت عنوان simrank نگاه می کنیم که این مفهوم به معنی کشف شباهت بین گره های یک گراف است.

ما شمارش مثلثی را به عنوان روشی برای سنجش ارتباط در یک جامعه استفاده می کنیم.

ما برای اندازه گیری دقیق و تقریبی اندازه مجموعه ها و گره های در یک گراف ، الگوریتم های مناسبی را ارائه می دهیم.

سرانجام ، برای محاسبه بستار انتقالی ، به الگوریتم های ویژه ای را بررسی می کنیم.

10.1 شبکه های اجتماعی به صورت گراف

بحث خود را در مورد شبکه های اجتماعی با معرفی یک مدل گرافی آغاز می کنیم.

هر نمودار گرافی مناسب به نمایش یک اجتماع در شبکه های اجتماعی نیست.

بنابراین، ما درباره ی ایده ی اصل محلیت که جز ویژگی های اصلی شبکه های اجتماعی است با کمک نودها و یال ها در گراف ها صحبت می کنیم. با کمک گره ها و یال ها تمایل خوشه بندی در شبکه ها بررسی می کنیم.

در این بخش همچنین برخی از انواع مختلف شبکه های اجتماعی که در عمل مورد استفاده قرار می گیرند را بررسی می کنیم.

10.1.1 شبکه اجتماعی چیست؟

وقتی به یک مفهوم شبکه اجتماعی فکر می کنیم ، به فیس بوک ، توییتر ، Google+ یا وب سایت دیگری فکر می کنیم که "شبکه اجتماعی" نامیده می شود و در واقع این شبکه ها نماینده ای از مفهوم شبکه های اجتماعی هستند.

ویژگی های اساسی یک شبکه اجتماعی عبارتند از:

1. مجموعه ای از اشخاص در شبکه های اجتماعی در شبکه وجود دارند که به طور معمول این موجودات مردم هستند اما می توانند چیزهای دیگری نیز باشند؛ در بخش 10.1.3 به مثالهای بیشتری در این مورد می پردازیم.
2. حداقل یک رابطه بین موجودیت های موجود (کاربران) در شبکه های اجتماعی وچود دارد. به رابطه ی بین موجودیت ها در فیس بوک Relationship می گویند. ارتباط ها یا وجود داشته یا ندارد پس دو نفر یا دوست هستند یا نیستند؛  
   گرچه در گراف های شبکه های اجتماعی رابطه ها دارای یک درجه هستند.  
   این درجه ها می تواند مقدار گسسته داشته باشد. به عنوان مثال در گوگل پلاس این درجه ها با عنوان های دوستان، خانواده، آشنایان و ... شناخته می شود.  
   این درجه می تواند یک عدد صحیح باشد یا یک عدد کسری باشد که از میزان صحبت بین دونفر بدست می آید.
3. محلیت در این شبکه ها به صورت غیرتصادفی است. این شرط برای نرمال سازی سخت ترین شرط است. اما ارتباطش به گرایش خوشه ها و روابط آنها می پردازد. یعنی اگر موجودیت A به هر دو موجودیت B و C مربوط باشد احتمال رخداد آن از میانگین B و C بدست می آید.

10.1.2 شبکه های اجتماعی به صورت گراف

شبکه های اجتماعی به طور معمول به عنوان گراف هایی مدل سازی می شوند که بعضا از آنها به عنوان یک گراف اجتماعی یاد می کنیم.

در این نمودار گراف گره ها موجودیت ها هستند و یال ها اتصال و ارتباط بین دو موجودیت (کاربر) هستند. اگر در این گراف درجه ای موجود باشد روی یال ها برچسب زده می شود.

غالبا گراف های موجود در شبکه های اجتماعی بدون جهت هستند؛ مثل گراف دوستان در شبکه اجتماعی فیس بوک.

اما می توان گراف های جهت دار هم داشت؛ مانند نمودار فالورهای توییتر یا گوگل پلاس.

مثال 10.1 : شکل 10.1 نمونه ای از یک گراف کوچک شبکه اجتماعی است؛ گره های این گراف از A تا G نامگذاری شده است. رابطه ی موجود در این گراف به عنوان دوست شناخته می شود که یال ها را تشکیل می دهد.

به عنوان مثال موجودیت B با A و C و D دوست می باشد.

آیا این گراف واقعا نشان دهنده ی یک شبکه ی اجتماعی و نمایش روابط آنها است؟

ابتدا توجه داشته باشید که گراف زیر دارای 9 یال می باشد.



شکل 10.1 : نمونه ای از شبکه اجتماعی کوچک

. در واقع 21 جفت گره می تواند در این شبکه دارای لبه باشد یا حداکثر یال های موجود در این شبکه 21 باشد.

فرض کنید X و Y و Z نودهای شکل 10.1 هستند که بین x و y و همچنین بین x و z دارای یال می باشد.

احتمال وجود یال بین y و Z چقدر است؟

اگر این نمودار بزرگ باشد این احتمالا به صورت کسری می باشد. یعنی در واقع 9 تقسیم بر 21 که برابر 0.429 می شود.

اما ، از آنجا که نمودار كوچك است ، بین احتمال واقعی و نسبت تعداد لبه ها به تعداد جفت گره ها تفاوت قابل ملاحظه ای وجود دارد.

از آنجایی که می دانیم یال های (X، Y) و (X، Z) وجود دارد ، تنها 7 یال باقی مانده است. این 7 یال می توانند بین هر 19 گره باقی مانده از گره ها قرار بگیرند.

بنابراین احتمال یک لبه بین z و y برابر 7 تقسیم بر 19 که برابر 0.368 می شود است.

حال باید این احتمال را محاسبه کنیم که لبه (Y ، Z) در شکل 10.1 وجود داشته باشد ، با توجه به اینکه لبه ها (X ، Y) و (X ، Z) وجود دارند. آنچه در واقع باید حساب کنیم ، جفت گره هایی است که می توانند Y و Z باشند ، بدون اینکه نگرانی در مورد کدام گره Y باشد و کدام Z باشد. اگر X A باشد ، باید Y و Z به ترتیب B و C باشند. از آنجا که لبه (B ، C) وجود دارد ، A یک مثال مثبت (که در آن لبه وجود دارد) است و هیچ نمونه منفی (جایی که لبه وجود ندارد) در آن نیست.

در مواردی که X به جای C ، E یا G است نتایج یکسان است.

در هر حالت ، X فقط دو همسایه دارد و لبه بین همسایگان وجود دارد. بنابراین ، ما تاکنون چهار نمونه مثبت و صفر مثال منفی را دیده ایم.

حال ، در نظر بگیرید که X = F. F دارای سه همسایه ، D ، E و G است. لبه هایی بین دو سه جفت همسایه وجود دارد ، اما هیچ لبه ای بین G و E وجود ندارد. بنابراین ، ما دو نمونه مثبت دیگر را می بینیم و اولین نمونه منفی خود را می بینیم.

اگر X = B ، دوباره سه همسایه وجود دارد ، اما فقط یک جفت همسایه ، A و C ، یک لبه دارند. بنابراین ، ما دو مثال منفی دیگر ، و یک مثال مثبت ، برای کل هفت مثبت و سه منفی داریم.سرانجام ، وقتی X = D ، چهار همسایه وجود دارد. از شش جفت همسایه ، فقط دو نفر بین آنها لبه دارند.

بنابراین ، تعداد کل نمونه های مثبت 9 و تعداد کل نمونه های منفی 7 است. در شکل 10.1 می بینیم که کسر ما برابر 9 تقسیم بر 16 است که در واقع برابر 0.563 می شود. این کسر خیلی بیشتر از مقدار قابل انتظار ما که 0.368 است می باشد.

در نتیجه شکل 10.1 واقعا اصل locality در شبکه های اجتماعی را نشان می دهد.

10.1.3 انواع شبکه های اجتماعی

نمونه های زیادی از شبکه های اجتماعی وجود دارد که ماهیت دوستان ندارد. در اینجا ، اجازه دهید تعدادی از نمونه های دیگر شبکه های اجتماعی که با اصل locality روابط را نشان می دهند ذکر کنیم.

شبکه های تلفن

در اینجا گره ها شماره تلفن ها را نشان می دهند ، که در واقع افراد هستند. اگر در طی مدت زمان مشخصی مانند ماه گذشته یا از ابتدا تا به حال بین این تلفن ها تماس برقرار شده باشد بین دو گره وجود دارد. یال ها را می توان با تعداد تماس های انجام شده بین این تلفن ها در طول دوره ی مشخص وزن داد. جوامع در یک شبکه تلفنی از گروههایی تشکیل می شوند که مرتباً ارتباط برقرار می کنند: برای مثال گروه هایی از دوستان ، اعضای یک باشگاه یا افرادی که در همان شرکت کار می کنند.

شبکه های ایمیلی

در شبکه های ایمیلی گره ها آدرس ایمیل افراد را نشان می دهند. یال ها بیانگر وجود حداقل یک ایمیل بین دو آدرس ایمیل می باشد. از طرف دیگر ممکن است لبه ها در این نوع شبکه ها به صورت یک طرفه یا دوطرفه باشند. از نمایش هرزنامه به عنوان دوست در این شبکه ها خودداری می شود یا با رویکردی آدرس های ایمیل هرزنامه را با لبه های ضعیف و لبه های دیگر را با لبه های قوی تر نشان می دهیم.

اجتماعاتی که در شبکه های ایمیلی مشاهده می شود از همان گروه بندی هایی هستند که در ارتباط با شبکه های تلفنی از آنها یاد کردیم. یک نوع دیگر برای مرتب سازی شبکه های ایمیلی از افرادی که از طریق تلفن های همراه متن های خود را می نویسند است.

شبکه های همکاری

گره ها افرادی را نشان می دهند که مقالات تحقیقاتی را منتشر کرده اند. بین دو فرد که یک یا چند مقاله را به طور مشترک منتشر کرده اند ، لبه وجود دارد. به صورت اختیاری ، می توان لبه ها را با تعداد انتشارات مشترک برچسب گذاری کرد. گروه های این نوع شبکه نویسنده هایی هستند که روی یک موضوع خاص کار می کنند. نمای جایگزین از همان داده ها به عنوان گرافی است که در آن گره ها مقالات هستند. اگر حداقل یک نویسنده مشترک داشته باشند ، دو مقاله به یک لبه متصل می شوند. اکنون ، گروه هایی را تشکیل می دهیم که مجموعه ای از مقالات در همین موضوع را شامل هستند.

چندین نوع داده دیگر وجود دارد که دو شبکه را به روشی مشابه ایجاد می کند.

به عنوان مثال ، می توانیم به افرادی که مقالات ویکی پدیا را برای بار اول منتشر می کنند و افرادی که مقاله هایی را ویرایش می کنند تقسیم کنیم. اگر ویرایش یک مقاله به صورت مشترک باشد ، دو گره ویرایشگر به یکدیگر متصل هستند. گروه هایی از ویراستاران که به صورت مشترک کار کرده اند را در یک دسته قرار می دهیم. بطور مضاعف ، می توانیم شبکه ای از مقالات بسازیم و در صورت ویرایش آنها توسط همان شخص ، مقاله ها را وصل کنیم.

در اینجا ، ما مقالاتی را درمورد موضوعات مشابه یا مرتبط با هم جمع می کنیم.

در واقع ، داده های مربوط به همکاری ، همانطور که در فصل 9 مورد بحث قرار گرفت ، اغلب می توانند به عنوان تشکیل یک جفت شبکه ، یکی برای مشتریان و دیگری برای محصولات مشاهده شوند.

مشتریانی که کالاهای مشابهی را خریداری می کنند ، به عنوان مثال ، کتاب های علمی تأیید می کنند ، جوامع تشکیل می دهند و بصورت دوگانه ، کالاهایی که توسط همان مشتریان خریداری می شوند ، جوامع را تشکیل می دهند؛

به عنوان مثال ، تمام کتاب های علمی تولیدی.

مثال های دیگر از گراف های شبکه های اجتماعی

بسیاری از پدیده های دیگر گراف هایی ا ایجاد می کنند که چیزی شبیه به گراف های شبکه های اجتماعی است ، به خصوص نمایش محلیت ها.

مثالهای این بخش عبارتند از: شبکه های اطلاعاتی (اسناد ، گراف های در بستر وب ، ثبت اختراعات) ، شبکه های زیرساختی (جاده ها ، هواپیماها ، لوله های آب ، نیروگاهها) ، شبکه های بیولوژیکی (ژن ها ، پروتئین ها ، شبکه های غذایی حیوانات که یکدیگر را می خورند) و همچنین انواع دیگر ، مانند شبکه های خرید محصول (به عنوان مثال ، Groupon).

10.1.4 گراف ها دارای چندین نوع گره

پدیده های اجتماعی دیگری وجود دارند که موجودیت های مختلفی را درگیر می کنند. تحت عنوان "شبکه های همکاری" ، انواع مختلفی از گراف ها که از دو نوع گره واقعاً تشکیل شده اند را دیده ایم.

شبکه های نویسندگی می توانند گره های نویسنده و گره های کاغذ را تشکیل دهند. در بحث بالا ، ما با از بین بردن گره های یکی از این دو نوع ، دو شبکه اجتماعی ایجاد کردیم اما لازم نیست این کار را انجام دهیم. ما می توانیم به طور کلی به ساختار فکر کنیم.

به عنوان مثال پیچیده تر ، کاربران در سایتی مانند del.icio.us برچسب ها را در صفحات وب قرار می دهند. بنابراین سه نوع مختلف موجودیت وجود دارد: کاربران ، برچسب ها و صفحات. ممکن است فکر کنیم اگر تمایل به استفاده از همان برچسب ها به طور مکرر یا اگر تمایل به برچسب زدن به همان صفحات دارند ، کاربران به نوعی به یکدیگر متصل شده اند. به طور مشابه ، برچسب ها اگر در همان صفحات ظاهر شوند می توانند مرتبط باشند؛ یا اگر برچسب ها توسط کاربران زیادی استفاده شود ممکن است صفحات مشابه به نظر برسند.

یک روش طبیعی برای نمایش چنین اطلاعاتی است که به عنوان گراف k-partite شناخته می شود؛ که k همیشه بزرگتر از 1 است. در بهش 8.3 گرافی دارای k=2 را مشاهده کرده ایم. به طور کلی یک نمودار k-partite از مجموعه گره هایی با k جداکننده تشکیل شده است که بین گره های یک گروه هیچ یالی وجود ندارد.

مثال 10.2 : شکل 10.2 نمونه ای از یک گراف k-partite است که مقدار k = 3 است. سه نوع گره وجود دارد: گره کاربران را در این گراف با {U1, U2} و گره برچسب ها را با {T1, T2, T3, T4} و گره صفخات وب را با {W1, W2, W3} نمایش می دهیم. توجه کنید تمام یال ها بین دو مجموعه مختلف متصل هستند.

ممکن است حس کنید این نمودار اطلاعاتی در مورد سه نوع موجودیت را نشان می دهد. به عنوان مثال لبه ی (U1, T2) به این معنی است که کاربر U1 برچسب T2 را حداقل در یک صفحه ی وب قرار داده است.

توجه کنید که گراف جزئیاتی را که می تواند مهم باشد را به ما نمی گوید. به عنوان مثال برای نمایش اینکه چه کسی چنین برچسبی را در این صفحه قرار داده است نیاز به نمایش پیچیده تری مانند روابط موجود در بانک های اطلاعاتی سه ستونه نیاز داریم.



شکل 10.2 : نمودار k-partite با مقدار k=3 برای نمایش کاربران، برچسب ها و صفحات وب

10.1.5 تمرینات مربوط به بخش 10.1

تمرین 10.1. : لبه های گراف G را به عنوان گره های G’ در نظر می گیریم.

1. اگر (X ، Y) یالی از گراف G باشد ، XY ، نمایانگر مجموعه ای هماهنگ از X و Y گره ای از گراف G’ است. توجه داشته باشید که XY و Y X یک گره G’ را نشان می دهند ، و نه دو گره مختلف.
2. اگر (X ، Y) و (X ، Z) لبه های گراف G باشند ، در گراف G’ یک لبه بین XY و XZ وجود دارد. اگر گره های G’ که این گره ها نمایان هستند دارای یک گره (از G) مشترک هستند ، گره های G’ بین آنها یال دارد.
3. اگر ساخت گراف را به صورت دوتایی را در شبکه ای از دوستان بکار گیریم ، تعبیر لبه های نمودار نتیجه چیست؟
4. ساخت دوتایی را در گراف شکل 10.1 اعمال کنید.
5. درجه گره XY در گراف G’ چگونه با درجه X و Y در گراف G ارتباط دارد؟
6. تعداد لبه های گراف G’ مربوط به درجه گره های گراف G توسط یک فرمول خاص است. آن فرمول را کشف کنید.
7. آنچه ما آنرا به عنوان دوگان نامیدیم در واقع یک دوتایی واقعی نیست؛ زیرا استفاده از این روش ساختن گراف در گراف G’ لزوماً یک ایزومورف گراف G می دهد. یک نمونه از گراف G را بدهید که در آن دوتایی از G’ از نظر گراف G ایزومورف است و مثال دیگری هم بدهید که در آن گرافG’ از نظر G ایزومورف نیست.

10.2 خوشه بندی گراف شبکه های اجتماعی

جنبه مهم شبکه های اجتماعی این است که آنها حاوی جوامع موجوداتی هستند که توسط بسیاری از لبه ها به هم وصل می شوند. به عنوان مثال ، اینها با گروهی از دوستان در مدرسه یا گروههایی از محققان علاقمند به همان موضوع مطابقت دارد. در این بخش ، خوشه بندی نمودار را به عنوان راهی برای شناسایی جوامع در نظر می گیریم. به نظر می رسد که تکنیک هایی که در فصل 7 آموخته ایم ، معمولاً برای مشکل خوشه بندی نمودارهای شبکه های اجتماعی نامناسب است.

10.2.1 معیار فاصله در گراف شبکه های اجتماعی

اگر بخواهیم از تکنیک های خوشه بندی استاندارد در یک گراف شبکه های اجتماعی استفاده کنیم ، اولین گام ما تعیین یک روش اندازه گیری فاصله است. هنگامی که یال های نمودار دارای برچسب هستند ، این برچسب ها بسته به آنچه که آنها نشان می دهند ، می توانند به عنوان اندازه گیری فاصله قابل استفاده باشند. اما هنگامی که لبه ها بدون برچسب هستند ، مانند نمودار «دوستان» ، برای تعیین فاصله ی مناسب ، کار زیادی نمی توان انجام داد.

اولین فرض ما این است که در نظر بگیریم گره ها نزدیک هستند و اگر یالی بین آنها باید یا نباشد دارای یک فاصله ی معین است. بنابراین ، می توان گفت که فاصله d(x,y) در صورت وجود لبه (x ، y) 1 است و در صورت عدم وجود چنین لبه ای 0 است. ما می توانیم از دو مقدار دیگری مثل 1 و ∞ استفاده کنیم تا زمانی که فاصله لبه ها به هم نزدیک تر باشد.

هیچ یک از معیارهای بررسی فاصله با دو ارزش 0 و 1 یا 0 و ∞ یک معیار واقعی و درست برای فاصله نیست. دلیل این امر این است که در هنگام اتصال مثلثی بین گره ها این مقدار فاصله بین دو گره نقض می شود. یعنی اگر لبه های (A ، B) و (B ، C) وجود داشته باشد ، اما هیچ لبه ای (A ، C) وجود ندارد ، در این صورت فاصله از A تا C از مجموع مسافت های A تا B از C بیشتر می شود. می توانیم برای این مشکل فاصله ی تا یک لبه را به طور مستقیم 1 در نظر بگیریم و برای missing edge ها از فاصله ی 1.5 استفاده کنیم که اینکار مشکل بالا را حل می کند اما مشکل توابع محاسبه ی فاصله به نابرابری مثلثی محدود نمی شود که در بخش بعدی این موضوع را خواهیم دید.

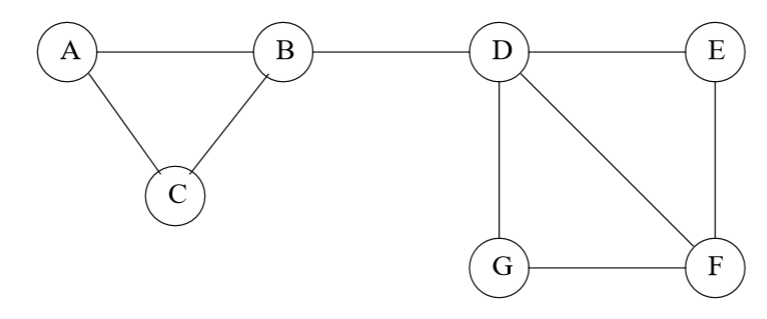
10.2.2 اعمال متدهای خوشه بندی استاندارد

از بخش 7.1.2 به یاد بیاورید که دو رویکرد کلی برای خوشه بندی وجود دارد: سلسله مراتبی (agglomerative) و تعیین امتیاز. در اینجا قصد این را داریم که چگونگی کار هر یک از این موارد را روی گراف های شبکه های اجتماعی بررسی کنیم.

ابتدا روش های سلسله مراتبی را که در بخش 7.2 است را در نظر بگیرید. به طور خاص ، فرض کنید که ما به عنوان فاصله متقاطع از حداقل فاصله بین گره های دو خوشه استفاده می کنیم.

خوشه بندی سلسله مراتبی از یک گراف شبکه های اجتماعی با ترکیب دو گره که به یک لبه متصل هستند ، آغاز می شود. به طور موفقیت آمیز ، لبه هایی که بین دو گره از یک خوشه یکسان نیستند به طور تصادفی انتخاب می شوند تا خوشه هایی را که دو گره آنها به آنها تعلق دارد ، ترکیب کنند. گزینه ها تصادفی هستند ، زیرا تمام مسافت هایی که توسط یک لبه نشان داده می شوند یکسان هستند.

مثال 10.3 : نمودار 10.1 در اینجا مانند شکل 10.3 تکرار شده است. اول ، بگذارید درمورد اینکه مجموعه ها چیست ، توافق کنیم. در بالاترین سطح ، به نظر می رسد که دو اجتماع {A,B,C} و {D,E,F,G} وجود دارند. با این حال ، ما همچنین می توانیم {D ، E ، F} و {D ، F ، G} را به عنوان دو زیرمجموعه از {D,E,F,G} در نظر بگیریم. این دو زیر مجموعه در دو عضوشان با هم همپوشانی دارند و بنابراین هرگز نمی توان با یک الگوریتم خوشه بندی خالص آنها را شناسایی کرد. سرانجام ، ما می توانیم هر جفت از افراد را که به یک لبه متصل هستند ، به عنوان یک جامعه با اندازه 2 در نظر بگیریم ، هرچند که چنین اجتماعاتی نگران کننده نیستند.



شکل 10.3 : تکرار مجدد شکل 10.1

مشکل خوشه بندی سلسله مراتبی از یک گراف مانند شکل 10.3 این است که در بعضی از نقاط احتمالاً ما ترکیب B و D را انتخاب می کنیم ، حتی اگر آنها مطمئناً در خوشه های مختلف قرار داشته باشند. دلیل اینکه ما احتمالاً B و D را با هم ترکیب کرده ایم این است که D و هر خوشه ای که حاوی آن باشد به همان اندازه نزدیک به B و هر خوشه ای است که حاوی آن باشد ، همانطور که A و C به B وجود دارد. حتی یک احتمال 1/9 وجود دارد که اولین کاری که ما انجام می دهیم ترکیب B و D در یک خوشه است.

مواردی وجود دارد که می توانیم برای کاهش احتمال خطا انجام دهیم. ما می توانیم چندین بار خوشه بندی سلسله مراتبی را اجرا کنیم و اجرا را انتخاب کنیم که منسجم ترین خوشه ها را داشته باشد. ما می توانیم از روش پیچیده تری برای اندازه گیری فاصله بین خوشه های بیش از یک گره استفاده کنیم ، همانطور که در بخش 7.2.3 بحث شده است. اما مهم نیست که چه کاری انجام می دهیم ، در یک نمودار بزرگ با بسیاری از جوامع ، شانس قابل توجهی وجود دارد که در مراحل اولیه باید از برخی لبه ها استفاده کنیم که دو گره را به هم متصل می کنند که در هیچ جامعه بزرگی وجود ندارند.

حال یک روش تعیین تکلیف را در مورد خوشه بندی شبکه های اجتماعی در نظر بگیرید. باز هم ، این واقعیت که همه لبه ها در یک فاصله قرار دارند ، تعدادی از عوامل تصادفی را معرفی می کند که منجر به اختصاص برخی گره ها به خوشه اشتباه می شود. با یک مثال باید نکته را نشان دهد.

مثال 10.4 : فرض کنید ما روش k-means را عنوان خوشه بندی در نظر بگیریم. در این خوشه بندی k=2 می باشد. اگر انتخاب دو گره را به صورت تصادفی انجام دهیم ممکن است هر دو در یک خوشه قرار بگیرند. اگر مانند بخش 7.32 با یک گره به صورت تصادفی شروع کنیم و بعد گره دیگری را انتخاب کنیم نتجه را خیلی بهتر نکرده ایم. از این طریق می توانیم هر جفت گره ای را که به یک لبه وصل نشده است ، انتخاب کنیم ، به عنوان مثال ، E و G که در شکل 10.3 است. با این حال فرض کنید ما با دو گره مناسب مثل B و F شروع کنیم. سپس A و C را به خوشه B اختصاص می دهیم و E و G را به خوشه F اختصاص می دهیم. اما گره D به اندازه ای که متعلق به F می باشد به همان اندازه متعلق به B است ، بنابراین می تواند به هر شکلی پیش برود ، حتی اگر "آشکار" باشد که D متعلق به F است باز هم این اتفاق می افتد.

اگر تصمیمی در مورد محل قرارگیری D به تعویق بیفتد تا زمانی که گره های دیگری را به خوشه ها اختصاص ندهیم ، احتمالاً تصمیم درست می گیریم. به عنوان مثال ، اگر ما یک گره را با کمترین مقاومت متوسط به همه گره های خوشه ، به خوشه اختصاص دهیم ، باید D را به خوشه F اختصاص دهیم ، مادامی که سعی نکنیم D را قبل از اینکه گره های دیگری اختصاص دهند ، قرار می دهیم. با این حال ، در نمودارهای بزرگ ، مطمئناً در برخی از اولین گره هایی که قرار می دهیم اشتباه می کنیم.

10.2.3 مفهوم Betweenness

از آنجا که در روش های استاندارد خوشه بندی مشکل خوشه بندی اختصاصی برای اجتماعات گراف شبکه های اجتماعی وجود دارد در این بخش یکی از ساده ترین روش ها را که بر اساس تعیین حاشیه هایی که احتمالاً در یک جامعه قرار دارند ، در نظر می گیریم.

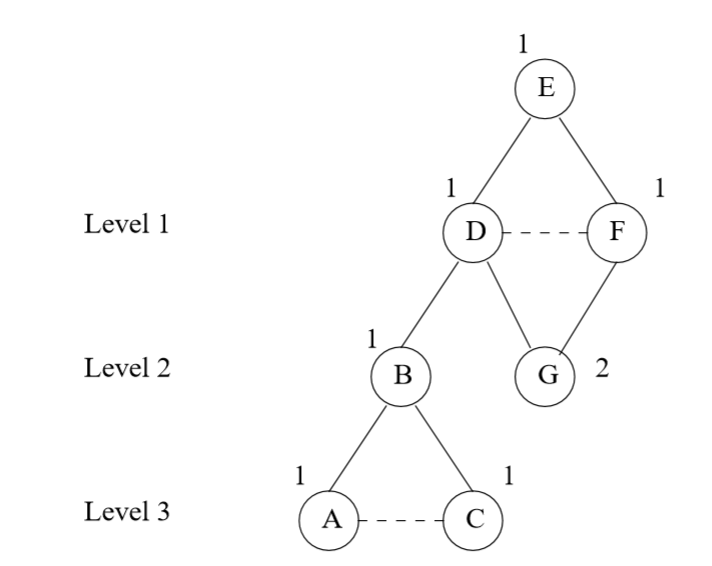
باید اتصال لبه ی (a ، b) به تعداد جفت گره های x و y به گونه ای باشد که لبه (a ، b) در کوتاهترین مسیر بین x و y قرار داشته باشد. به بیان دقیق تر ، از آنجا که می توانید چندین مسیر کوتاه بین x و y وجود داشته باشد ، لبه (a ، b) با کسری از آن کوتاه ترین مسیرها که شامل لبه (a ، b) است ، اعتبار دارد. مانند گلف ، که در آن امتیاز بالا بد است. این نشان می دهد که لبه (a ، b) بین دو اجتماع مختلف جریان دارد. یعنی ، a و b متعلق به یک مجموعه نیستند.

مثال 10.5 : در شکل 10.3 لبه (B ، D) بالاترین حد فاصل را دارد؛ در حقیقت ، این لبه در هر کوتاه ترین مسیری بین هر یک از A ، B و C به هر یک از D ، E ، F و G قرار دارد. فاصله آن 12 است. در مقابل ، لبه (D ، F) فقط در چهار مسیر کوتاه قرار دارد: آنهایی که از A ، B ، C و D تا F هستند.

10.2.4 الگوریتم Girvan-Newman

برای بهره برداری از لبه ها ، باید تعداد کوتاهترین مسیری که در هر لبه طی می شود را محاسبه کنیم. روشی را به نام الگوریتم Girvan-Newman (GN) توصیف می کنیم ، که یک بار از هر گره X بازدید می کند و تعداد کوتاهترین مسیرها را از X به هر گره دیگری که از هر یک از لبه ها عبور می کند محاسبه می کند. این الگوریتم با روش (BFS) در گراف ، از گره X شروع می کند. توجه داشته باشید که سطح هر گره در نمایش BFS طول کوتاهترین مسیر از X تا آن گره است. بنابراین ، لبه هایی که بین گره ها در یک سطح قرار دارند هرگز نمی توانند بخشی از کوتاهترین مسیری از X باشند.

لبه های بین سطوح ، لبه های DAG نامیده می شوند ("DAG" مخفف directed, acyclic graph است). هر لبه DAG بخشی از حداقل یک مسیر کوتاه از ریشه X خواهد بود. اگر یک لبه ی Dag در (Y,Z) وجود داشته باشد جایی که Y در سطح بالاتر از Z قرار دارد (یعنی نزدیک به ریشه) ، آنگاه ما Y را والدین Z و Z فرزند Y می نامیم ، گرچه آنها لزوماً در DAG والدین یکتایی نیستند که به عنوان یک درخت قرار بگیرند.



شکل 10.4 : مرحله اول الگوریتم Girvan-Newman

مثال 10.6 : شکل 10.4 روش breadth-ﬁrst از گراف شکل 10.3 است که از گره E شروع می شود. یال های غیرخط چین لبه های DAG هستند و لبه های خط چین گره ها را در همان سطح قرار می دهند.

مرحله دوم الگوریتم Girvan-Newman ، برچسب زدن هر گره با تعداد کوتاهترین مسیری است که از ریشه به آن می رسد. با برچسب زدن به ریشه شروع کنید. سپس ، از بالا به پایین ، هر گره Y را بر اساس برچسب های والدین آن ، برچسب گذاری کنید.

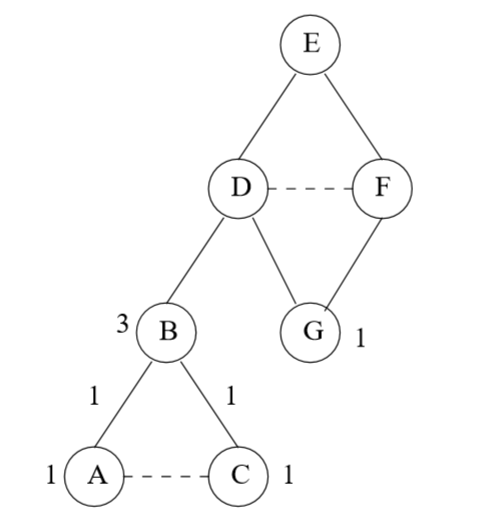
مثال 10.7 : در شکل 10.4 برچسب های هر یک از گره ها نشان داده شده است. ابتدا ریشه که E می باشد را با 1 علامت گذاری کنید. در سطح اول گره های D و F هستند. هر کدام فقط گره E را به عنوان والدین دارند ، بنابراین آنها نیز دارای برچسب 1 هستند. گره های B و G در سطح دو قرار دارند. گره B فقط والدین گره D را دارد ، بنابراین برچسب B برابر با برچسب D است که 1 است. با این حال ، گره G دارای والدین D و F است ، بنابراین برچسب آن مجموع برچسب های آنها یا 2 است. سرانجام ، در سطح سه، گره A و C هر یک فقط والدین B را دارند ، بنابراین برچسب های آنها برچسب B است که 1 است.

مرحله سوم و پایانی محاسبه برای هر لبه ی e و مجموع بیش از همه گره ها Y از کسری از کوتاه ترین مسیرها از ریشه X تا Y است که از e عبور می کند. این محاسبه شامل محاسبه این هزینه برای گره ها و لبه ها ، از پایین است. به هر گره به غیر از ریشه هزینه ی 1 داده می شود ، که کوتاه ترین مسیر را به آن گره نشان می دهد. این هزینه ممکن است بین گره ها و لبه های بالا تقسیم شود ، زیرا ممکن است چندین مسیر کوتاه مختلف به گره وجود داشته باشد. قوانین محاسبه به شرح زیر است:

1. برگ ها در DAG دارای هزینه ی 1 هستند.
2. گره های غیر برگ دارای هزینه ی 1 به علاوه جمع هزینه ی DAG نسبت به نودهای سطح پایینتر است.
3. یک لبه DAG و بخش ورودی به گره Z از سطح فوق با هزینه ی Z متناسب با کسری از کوتاهترین مسیرها از ریشه تا Z که از E عبور می کند ، داده می شود. به طور کلی ، والدین Z را به شکلY1 ، Y2 ، ... ، Yk در نظر بگیرید و بگذارید pi تعداد کوتاهترین مسیرها از ریشه تا Yi باشد. این عدد در مرحله دو محاسبه شده است و توسط برچسب ها در شکل 10.4 نشان داده شده است. سپس هزینه لبه (Yi، Z) برابر است.

پس از انجام محاسبه هزینه ی هر گره به عنوان ریشه ، اعتبارات مربوط به هر لبه را جمع می کنیم. سپس ، از آنجا که هر کوتاهترین مسیر دو بار کشف شده است - یک بار وقتی که هر یک از نقاط انتهایی آن ریشه دارد - باید اعتبار هر لبه را به 2 تقسیم کنیم.

مثال 10.8 : محاسبه هزینه را برای روش BFS که از شکل است10.4 انجام می دهیم. ما باید از سطح سه شروع کنیم و به سمت بالا پیش برویم. اول ، A و C ، که برگ می شوند ، اعتبار 1 را دریافت می کنند. هر یک از این گره ها فقط یک والد دارند ، بنابراین اعتبار آنها به ترتیب به لبه ها (B ، A) و (B ، C) داده می شود.



شکل 10.5 : مرحله ی نهایی الگوریتم Girvan-Newman

در سطح دو ، G یک برگ است ، بنابراین اعتبار 1 را دریافت می کند. B برگ نیست ، بنابراین اعتبار آن برابر با 1 به علاوه اعتبارات در لبه های DAG که از زیر آن وارد شده است ، می شود. از آنجا که هر دو این لبه دارای اعتبار 1 هستند ، اعتبار B برابر 3 است. به طور شهودی 3 این واقعیت را نشان می دهد که تمام کوتاهترین مسیرها از E به A ، B و C از B عبور می کنند. شکل 10.5 اعتبارات اختصاص یافته تاکنون را نشان می دهد.

اکنون ، اجازه دهید به سطح 1 برویم. B تنها یک والد دارد ، D ؛ بنابراین لبه (D ، B) اعتبار کل B را دریافت می کند ، که 3 است. با این حال ، G دو والدین دارد ، D و F.

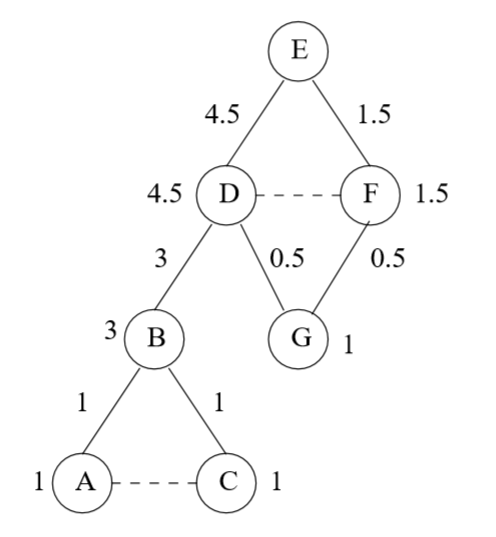
بنابراین ما باید اعتبار 1 را که G دارد بین لبه ها (D ، G) و (F ، G) تقسیم کنیم.

به چه نسبت تقسیم می کنیم؟

اگر برچسب های شکل 10.4 را بررسی می کنید ، می بینید که هر دو D و F دارای برچسب 1 هستند ، این نشان دهنده این واقعیت است که یک کوتاهترین مسیر از E به هر یک از این گره ها وجود دارد.

بنابراین ، ما نصف اعتبار G را به هر یک از این لبه ها می دهیم ؛ یعنی ، اعتبار آنها هر 1 / (1 + 1) = 0.5 است.

اگر برچسب های D و F در شکل 10.4 برابر 5 و 3 باشد ، به این معنی که کوتاه ترین مسیرها به D و فقط سه تا F وجود داشته است ، پس اعتبار لبه (D,G) برابر پنج هشتم است و لبه (F ، G) می توانست سه هشتم باشد.



شکل 10.6 : مرحله ی نهایی الگوریتم Girvan-Newman – با اتمام محاسبه ی اعتبارها(هزینه ها)

اکنون ، می توانیم اعتبارات را به گره ها در سطح 1 اختصاص دهیم. گره D اعتبار 1 را می گیرد به علاوه اعتبار لبه هایی که از زیر آن وارد می شوند ، 3 و 0.5 است. یعنی اعتبار گره D برابر 4.5 است. اعتبار F برابر 1 به علاوه اعتبار لبه (F، G) یا 1.5 است. سرانجام ، لبه های (E ، D) و (E ، F) به ترتیب اعتبار D و F را دریافت می کنند ، زیرا هر یک از این گره ها فقط یک والدین دارند. این اعتبارات همه در شکل 10.6 نشان داده شده است.

اعتبار هر یک از لبه ها در شکل 10.6 سهم در فاصله بین آن لبه به دلیل کوتاه ترین مسیرهای E. است. برای مثال ، این سهم برای لبه ی (E,D) برابر 4.5 است.

برای تکمیل محاسبه فاصله بین آنها ، باید این محاسبه را برای هر گره به عنوان ریشه تکرار کنیم و سهم ها را با هم جمع کنیم. سرانجام، باید 2 را تقسیم کنیم تا فاصله ی حقیقی بین آنها را پیدا کنیم ، زیرا هر کوتاهترین مسیر دو بار ، یک بار برای هر یک از نقاط پایانی آن کشف می شود.

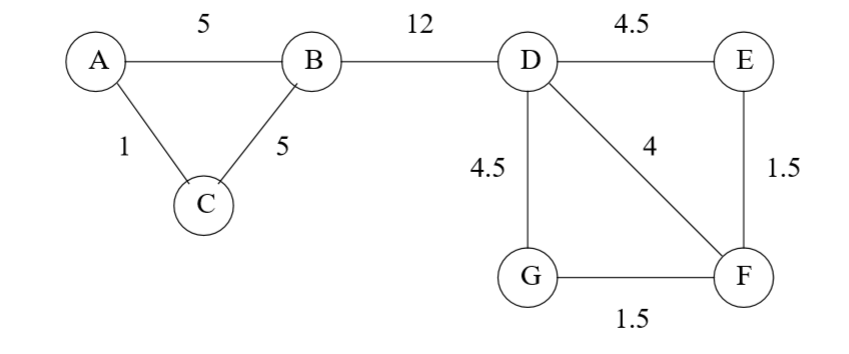
10.2.5 استفاده از مفهوم Betweenness در پیدا کردن مجموعه ها

هزینه بین لبه های نمودار چیزی شبیه به اندازه گیری مسافت بر روی گره های گراف است. این دقیقاً اندازه گیری مسافت نیست ، زیرا برای جفت گره هایی که به یک لبه وصل نشده اند ، تعریف نمی شود و ممکن است نابرابری مثلثی را حتی در صورت تعریف ، برآورده نکند. با این حال ، ما می توانیم با گرفتن لبه ها به منظور افزایش فاصله ، خوشه ای ایجاد کنیم و هر یک را یکبار به گراف اضافه کنیم. در هر مرحله ، اجزای متصل به گراف برخی خوشه ها را تشکیل می دهند. هرچه فاصله بین آنها بیشتر باشد ، لبه های بیشتری می گیریم و خوشه ها بزرگتر می شوند.

به طور معمول ، این ایده به عنوان فرآیند حذف لبه بیان می شود. با گراف و تمام یال های آن شروع کنید. سپس یال ها را با بیشترین فاصله حذف می کنیم، تا زمانی که نمودار به تعداد مناسبی از اجزای متصل شکسته شود.

مثال 10.9 : با مثال قبلی، گراف شکل 10.1 شروع می کنیم. ما این گراف را با فاصله بین هر لبه در شکل 10.7 می بینیم. محاسبه فاصله بین خواننده خواهد بود. تنها بخش دشوار شمارش این است كه مشاهده كنید كه بین E و G دو كوتاه ترین مسیر وجود دارد كه یكی از D عبور می كند و دیگری از طریق F.

بنابراین ، به هر یک از لبه های (D,E) و (E,F) و (D,G) و (G,F) با نیم کوتاه ترین مسیر اعتبار داده می شود.

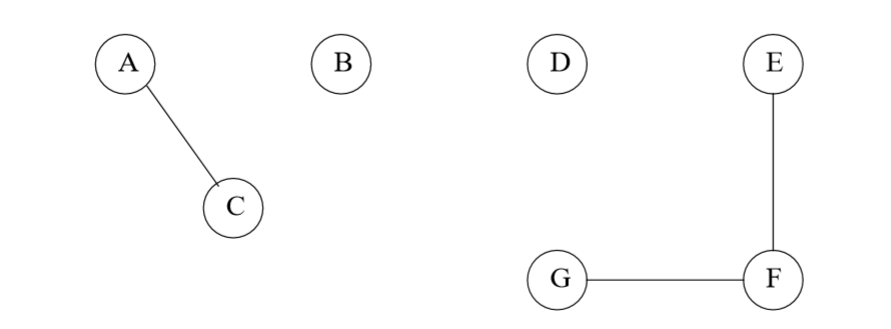


شکل 10.7 : امتیاز بین گره های گراف شکل 10.1

بدیهی است که لبه (B ، D) بالاترین حد فاصل را دارد ، بنابراین اولین بار برداشته می شود. این دقیقاً حس جامعه را به ما رهنمون می کند ، یعنی: {A ، B ، C} و {D ، E ، F ، G }.

با این حال ، ما می توانیم به حذف لبه ها ادامه دهیم. در مرحله بعد (A، B) و (B، C) با امتیاز 5 قرار دارند و پس از آن (D، E) و (D، G) با امتیاز 4.5.

سپس ، (D ، F) ، که نمره آن 4 است ، نمودار را ترک می کند. در شکل 10.8 گراف هایی را که می ماند می بینیم.



شکل 10.8 : تمام لبه های بین 4 یا بیشتر برداشته شده است.

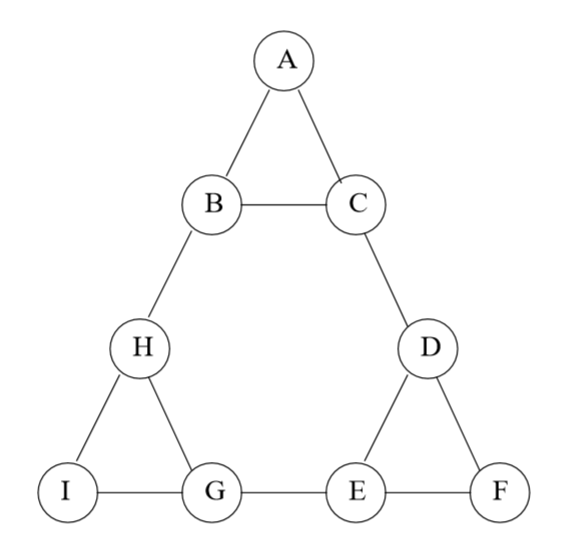
"اجتماعات" در شکل 10.8 عجیب به نظر می رسند. یکی از دلالت ها این است که A و C از نزدیک به یکدیگر گره خورده اند تا B. یعنی به نوعی B یک "خائن" جامعه است - A ، B ، C - زیرا او در خارج از جامعه دوست D دارد. به همین ترتیب ، D را می توان به عنوان "خائن" به گروه - {D,E,F,G} در نظر گرفت ، به همین دلیل در شکل 10.8 ، فقط E ، F و G به هم وصل می شوند.

|  |
| --- |
| سرعت بخشیدن در بحث محاسبه ی Betweenness  اگر ما روش بخش 10.2.4 را در یک نمودار از گره های n و لبه های e استفاده کنیم ، زمان محاسبه O (ne) زمان لازم برای محاسبه فاصله بین هر لبه است. یعنی BFS از یک گره واحد زمان O (e) را نیز می گیرد ، همانطور که دو مرحله برچسب زدن را انجام می دهید. باید از هر گره شروع کنیم ، بنابراین n از محاسبات شرح داده شده در بخش 10.2.4 وجود دارد.  اگر نمودار بزرگ باشد ما نمی توانیم نظمی را برای اجرای آن اجرا کنیم. با این وجود ، اگر زیرمجموعه ای از گره ها را بطور تصادفی انتخاب کنیم و از این ها به عنوان ریشه های جستجو برای اولین بار استفاده کنیم ، می توانیم به تقارن بین هر لبه ای که در اکثر برنامه ها خدمت می کند ، برسیم. |

10.2.6 تمرین های بخش 10.2

تمرین 10.2.1 : شکل 10.9 نمونه ای از گراف شبکه های اجتماعی است. از روش Girvan-Newman استفاده کنید تا تعداد کوتاهترین مسیرهای هر کدام از گره های زیر را که از هر یک از لبه ها عبور می کند ، تعیین کنید.

1. A
2. B



شکل 10.9 : گراف مربوط به تمرین

تمرین 10.2.2 : با استفاده از تقارن ، محاسبات تمرین 10.2.1 ، تمام موارد لازم برای محاسبه فاصله بین هر لبه است این محاسبات را انجام دهید.

تمرین 10.2.3: با استفاده از مقادیر فاصله از روش 10.2.2 ، نامزدهای معقول را برای جوامع در شکل 10.9 با حذف تمام لبه ها با فاصله بالای برخی از آستانه تعیین کنید.

10.3 کشف مستقیم مجموعه ها یا جوامع

در بخش قبلی ما با جدا کردن همه افراد در یک شبکه اجتماعی ، جوامع را جستجو کردیم. اگرچه این رویکرد نسبتاً مؤثر است ، اما محدودیت دارد. قرار دادن فرد در دو جامعه مختلف امکان پذیر نیست و همه به یک جامعه اختصاص داده می شوند. در این بخش ، با جستجوی زیرمجموعه هایی از گره ها که تعداد نسبتاً زیادی از لبه ها در بین آنها وجود دارد ، می توانیم تکنیکی برای کشف مستقیم جوامع مشاهده کنیم. جالب است که ، تکنیک انجام این جستجو بر روی یک گراف بزرگ ، شامل موارد سنگین اجرایی و به صورت مکرر است ، همانطور که در فصل 6 مورد بحث قرار گرفت.

10.3.1 پیدا کردن Cliques

اولین فکر ما در مورد اینکه چگونه می توانیم مجموعه های گره هایی که لبه های زیادی بین آنها وجود دارد را با ایجاد یک clique بزرگ (مجموعه ای از گره ها با لبه های بین هر دو آنها) شروع کنیم. با این حال ، این کار آسان نیست. نه تنها clique ها حداکثر NP کامل است ، بلکه به این معنا که حتی تقریب حداکثر clique سخت است.

علاوه بر این ، می توان مجموعه ای از گره ها را که تقریباً همه لبه ها در بین آنها وجود دارد در عین حال فقط دارای clique های نسبتاً کوچکی می باشد.

تمرین 10.10 : فرض کنید گراف ما دارای گره هایی به شماره 1،2 ، ... ، n باشد و یک لبه بین دو گره i و j وجود دارد ، مگر اینکه i و j باقیمانده با k تقسیم شوند. سپس کسری از لبه های ممکن که در واقع وجود داشته باشد تقریباً (k-1) / k است. تعداد زیادی clique با اندازه k وجود دارد که از آن {1,2,...,k} فقط یک clique است.

با این حال هیچ cliqueیی بزرگتر از k وجود ندارد. برای دیدن دلیل این موضوع ، مشاهده کنید که هر مجموعه گره k + 1 دارای دو عدد است که هنگام تقسیم بر k ، باقیمانده باقی می مانند.

این نکته کاربردی از "اصل لانه کبوتری" است. از آنجا که فقط باقیمانده های مختلف K ممکن است ، ما نمی توانیم بقایای مشخصی برای هر یک از گره های k + 1 داشته باشیم. بنابراین ، هیچ مجموعه ای از گره های k + 1 نمی تواند یک clique در این گراف باشد.

10.3.2 گراف های دوبخشی کامل

بحث گراف های دو طرفه را از بخش 8.3 به یاد بیاورید. یک گراف کامل دو طرفه از گره های s در یک طرف و گره های t در طرف دیگر تشکیل شده است که تمام لبه های ممکن بین گره های یک طرف و طرف دیگر موجود است. ما این گراف را توسط نشان می دهیم. شما باید قیاس بین گراف های کامل دو طرفه را به عنوان زیرگراف نمودارهای دو طرفه کلی و کلیشه ها به عنوان زیرگرافهای نمودارهای کلی ترسیم کنید.

در حقیقت ، غالباً به کلیشه گره های s به عنوان نمودار کاملی گفته می شود و Ks را نشان می دهد ، در حالی که یک زیرگراف کامل دو طرفه گاهی اوقات یک biclique خوانده می شود.

در حالی که همانطور که در مثال 10.10 دیدیم ، نمی توان تضمین کرد که یک گراف با بسیاری از لبه ها لزوماً دارای یک clique بزرگ است ، می توان تضمین کرد که یک نمودار دو طرفه با بسیاری از لبه ها دارای یک زیرگراف کامل دو طرفه کامل است.

ما می توانیم یک زیرگراف کامل دو طرفه (یا یک clique اگر یک مورد بزرگ را کشف کردیم) به عنوان هسته یک جامعه در نظر بگیریم و گره هایی با لبه های زیادی به اعضای موجود در جامعه اضافه کنیم.

اگر خود گراف همانطور که در بخش 10.1.4 مورد بحث قرار گرفته است k-partite باشد ، می توان گره هایی از دو نوع و لبه های بین آنها را در نظر گرفت تا یک گراف دو بخشی ایجاد کنیم. در این گراف دو بخشی ، ما می توانیم برای زیرگرافهای کاملاً دو طرفه به عنوان هسته جوامع جستجو کنیم.

به عنوان مثال ، در مثال 10.2 ، ما می توانیم بر روی برچسب ها و گره های صفحه از گرافیکی مانند شکل 10.2 تمرکز کنیم و سعی کنیم مجموعه های برچسب ها و صفحات وب را ایجاد کنیم.

چنین جامعه ای از برچسب ها و صفحات مربوطه تشکیل می شود که مستحق بسیاری از این برچسب ها هستند.

با این حال ، ما همچنین می توانیم از زیرگراف های کاملاً دو بخشی برای نمایش اجتماع در گراف های معمولی استفاده کنیم که گره ها همه از یک نوع هستند.

گره ها را بطور تصادفی به دو گروه مساوی تقسیم کنید.

اگر جامعه ای وجود داشته باشد ، انتظار داریم تقریباً نیمی از گره های آن در هر گروه قرار بگیرد و ما انتظار داریم که حدود نیمی از لبه های آن بین گروه ها قرار بگیرد.

بنابراین ، ما هنوز یک فرصت منطقی برای شناسایی یک زیرگراف کامل کامل دو طرفه در جامعه داریم.

به این هسته می توان گره هایی از هر دو گروه اضافه کرد ، در صورتی که دارای لبه هایی در بسیاری از گره هایی هستند که قبلاً متعلق به جامعه هستند.