د کابل پوهنتون دطبیعي علومو علمي ـــ څېړنیزه مجله
مجـــله علـــمی- تحقیقی علـــوم طبیعی پوهنتون کابل
Journal of Natural Sciences – Kabul University
e-ISSN:3078-6665 P-ISSN:3078-5561

کلسترسازی گراف

پوهنیار الیاس احمد احمدی[۱]، پوهنیار بسم الله دانش[۲]

[۲،۱] دیپارتمنت ریاضیات عمومی، پوهنځی ریاضیات، پوهنتون کابل، کابل، افغانستان

ایمیل: eliasahmad266@yahoo.com

چکیده

این مقاله به بررسی کلسترسازی گراف به عنوان یک مسأله بهینه‌سازی ترکیبی پرداخته و پیچیدگی محاسباتی الگوریتم‌های مرتبط را ارزیابی کرده است. هدف اصلی این تحقیق، تحلیل رابطه میان اهداف چند کلستری و اهداف مبتنی بر مودولاریتی و شناسایی محدودیت‌های مودولاریتی در کلسترسازی گراف‌ها می‌باشد. این تحقیق مروری-کیفی با استفاده از تحلیل‌های ریاضی و آزمایش الگوریتم‌های مختلف انجام شده است. یافته‌ها نشان داد که کلسترسازی گراف‌ها به دلیل فضای جستجوی نمایی یک مسأله NP - سخت بوده و مودولاریتی با وجود کاربرد گسترده، دچار خطای تفکیک است و قادر به شناسایی دقیق کلسترهای مقیاس کوچک نمی‌باشد. نتایج همچنین بیانگر برتری اهداف چند کلستری در شناسایی ساختارهای ظریف‌تر است. این تحقیق با تأیید یافته‌های پیشین و ارائه تحلیل‌های جدید، درک عمیق‌تری از بهینه‌سازی و کلسترسازی گراف‌ها فراهم کرده و مسیرهایی برای بهبود الگوریتم‌های موجود پیشنهاد داده است.

واژه‌های کلیدی: کلسترسازی گراف؛ بهینه‌سازی؛ تشخیص جامعه؛ پیچیدگی محاسباتی؛ تحلیل شبکه

Graph Clustering

Elias Ahmad Ahmadi[1], Besmillah Danish[2]

[1,2] Department of General Mathematics, Faculty of Mathematics, Kabul University, Kabul, Afghanistan

Email: eliasahmad266@yahoo.com

Abstract

This study examined graph clustering as a combinatorial optimization problem and evaluated the computational complexity of related algorithms. The main objective of this research was to analyze the relationship between multi-cluster objectives and modularity-based goals and to identify the limitations of modularity in graph clustering. This qualitative review study employed mathematical analyses and tests with various algorithms. The findings showed that graph clustering, due to its exponential search space, is an NP-hard problem, and modularity, despite its widespread use, suffers from a resolution limit and cannot accurately detect small-scale clusters. The results also highlighted the superiority of multi-cluster objectives in identifying more subtle structures. By confirming previous findings and offering new insights, this research provided a deeper understanding of optimization and graph clustering and suggested pathways to enhance existing algorithms.

Keywords: Graph Clustering; Optimization; Community Detection; Computational Complexity; Network Analysis

ارجاع: احمدی، ا. ا. و دانش، ب. (۱۴۰۳). کلسترسازی گراف. *مجله علمی- تحقیقی علوم طبیعی پوهنتون کابل*، *۷(٤)*، ۳۴۳-۳۷۰. https://doi.org/10.62810/jns.v7i4

مقدمه

گراف یک ساختار ریاضی است که ست‌ای از رأس‌ها و روابط جوره ای بین آن‌ها (اضلاع) را نمایش می‌دهد. اصطلاح "شبکه" تقریباً مترادف با اصطلاح گراف است و به طور خاص برای اشاره به یک سیستم پیچیده از عوامل متعامل استفاده می‌شود که در هسته خود می‌تواند با استفاده از یک گراف مودل‌سازی و مطالعه گردد. سیستم‌های پیچیده‌ای که توسط دانشمندان شبکه مورد مطالعه قرار می‌گیرند شامل شبکه‌های حمل و نقل (مانند شبکه‌های جاده‌ای و شبکه‌های هوایی)، شبکه‌های فناوری (مانند انترنت و شبکه‌های تیلفون)، شبکه‌های اجتماعی (مانند شبکه‌های دوستی) و شبکه‌های بیولوژیکی (مانند شبکه‌های تعامل پروتین، زنجیره‌های غذایی، سیستم‌های عصبی) هستند که تنها به چند نمونه اشاره شده است. یکی از بنیادی‌ترین مسائل در علم شبکه و نظریه گراف تشخیص ست‌هایی از رأس‌ها است که بیشتر به یکدیگر متصل هستند تا به بقیه گراف. این مسائل اساسی نام‌های مترادف بسیاری از جمله تقسیم‌بندی گراف، کلسترسازی گراف و تشخیص جامعه دارد.

تقسیم‌بندی یک گراف به کلسترهای متشکل از رأس‌های مرتبط، کاربردهای وسیع‌ای در زمینه‌های مختلف دارد. این روش در تشخیص جن‌های مرتبط در شبکه‌های بیولوژیکی بسیار مفید است، زیرا می‌تواند به درک بهتری از تعاملات جنی و مسیرهای بیولوژیکی کمک کند. همچنین، در پردازش تصویر و بینایی کامپیوتری، تقسیم‌بندی تصاویر به بخش‌هایی که با اشیاء موجود در تصویر تطابق دارند، کاربرد دارد. در تحلیل شبکه‌های اجتماعی، این روش برای تشخیص ساختار اجتماعی و گروپ‌ها و ارتباطات میان کاربران استفاده می‌شود. در زمینه محاسبات، تقسیم یک مسأله به بخش‌هایی که می‌توانند به طور مؤثرتر به صورت موازی پردازش شوند، در بهینه‌سازی عملکرد سیستم‌های محاسباتی و توزیع‌شده اهمیت زیادی دارد. همچنین، در طبابت، تشخیص نواحی و انواع مختلف بافت در اسکن‌های MRI به تشخیص دقیق‌تر و بهتر کمک می‌کند. در نهایت، این روش در تشخیص رفتارهای غیرعادی و فعالیت‌های تقلبی در شبکه‌های صحی و مالی کاربرد دارد و می‌تواند به جلوگیری از فعالیت‌های مشکوک و تقلبی کمک کند. این کاربردها نشان می‌دهند که کلسترسازی گراف ابزار قدرتمندی برای تحلیل ساختارهای پیچیده و کشف الگوها در دیتاهای مختلف است.

علاوه بر کاربردهای وسیع‌اش، کلسترسازی گراف یک مسأله بسیار غنی از نظر ریاضی است و از بُعد نظری به طور وسیع توسط ریاضی‌دانان، فزیک‌دانان، دانشمندان کامپیوتر، متخصصین احصاییه و کارشناسان از حوزه‌های مختلف علمی مورد مطالعه قرار گرفته است. با وجود نتایج بسیار و چشم‌انداز وسیع تخنیک‌های موجود، کلسترسازی گراف همچنان یک مشکل چالش‌برانگیز باقی مانده است.

اهداف این تحقیق به بررسی جامع و تحلیلی کلسترسازی گراف به عنوان یک مسأله بهینه‌سازی ترکیبی می‌پردازد. این تحقیق به دنبال ارزیابی پیچیدگی محاسباتی الگوریتم‌های کلسترسازی گراف و درک عمیق رابطه میان اهداف چند کلستری و اهداف مبتنی بر مودولاریتی است. همچنین، محدودیت‌های روش مودولاریتی در فرآیند کلسترسازی گراف‌ها مورد بررسی قرار گرفته و تلاش می‌شود تا نتایج معادل در کلسترسازی گراف‌ها شناسایی و تفسیر شوند. هدف نهایی، ارائه دیدگاهی روشن و ساختارمند از چالش‌ها و فرصت‌های بهینه‌سازی در این حوزه است.

با توجه به اهداف این تحقیق، پرسش‌های اصلی به شرح زیر مطرح می‌شوند:

۱. کلسترسازی گراف چگونه به عنوان یک مسأله بهینه‌سازی ترکیبی مطرح می‌شود؟

۲. پیچیدگی محاسباتی الگوریتم‌های کلسترسازی گراف چگونه ارزیابی می‌شود؟

۳. رابطه میان اهداف چند کلستری و اهداف مبتنی بر مودولاریتی چیست؟

۴. محدودیت‌های مودولاریتی در کلسترسازی گراف‌ها کدام‌هاهاند؟

۵ نتایج معادل در کلسترسازی گراف‌ها چه هستند و چگونه تفسیر می‌شوند؟

روش تحقیق

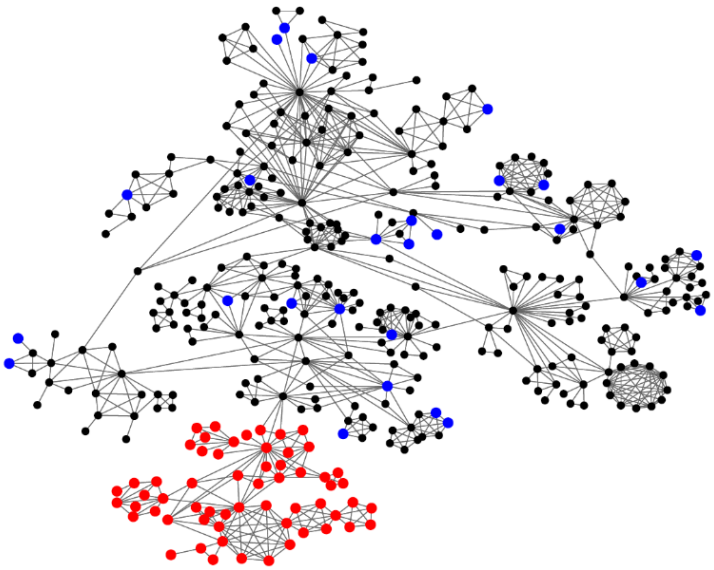
این تحقیق مروری کیفی می‌باشد که به تحلیل و ترکیب دیتاهای کیفی از مطالعات مختلف تمرکز دارد؛ برای انجام این تحقیق اولاً بعضی سایت‌های انترنتی؛ مانند گوگل سکالر، اسکاپَس، پاپمیت و غیره با استفاده از کلمات کلیدی همچو کلسترسازی گراف، بهینه‌سازی، تشخیص جامعه، پیچیده‌گی محاسباتی و تحلیل شبکه جستجو گردیده، مقالات و کتاب‌های که ارتباط نزدیک با عنوان این تحقیق داشتند انتخاب و به دقت مورد غور و بررسی قرار گرفتند. یاداشت‌ها و معلومات برداشته شده از این مقالات و کتاب‌ها که به صورت کیفی می‌باشد، در تکمیل این تحقیق مورد استفاده قرار گرفته است.

یافته‌های تحقیق

اصول کلسترسازی گراف

شکل ۱ تصویری از بزرگ‌ترین مؤلفه گراف متصل نیت-ساینس (Net-science) است. در این گراف رأس‌ها نشان‌دهنده محققان اکادمیک و اضلاع نشان‌دهنده هم‌نویسنده‌گی (نویسنده‌گی مشترک) در یک مقاله در باره شبکه ساینس هستند. دو ست از رأس‌ها در اینجا یکی به رنگ آبی و دیگری به رنگ سرخ برجسته شده‌اند. بدون نیاز به علم ریاضی، می‌توان به راحتی حدس زد که ست سرخ نظر به ست آبی بیشتر احتمال دارد تا به عنوان یک "جامعه" در نظر گرفته شود. رأس‌های آبی که به صورت تصادفی انتخاب شده‌اند، در سرتاسر گراف پراکنده بوده و ارتباطات کمی با یک‌دیگر دارند، در حالی

که رأس‌های سرخ به وضوح به هم متصل اند و عمدتاً از بقیه گراف جدا هستند. در واقع، رأس‌های سرخ نمایانگر یک جامعه از محققان اکادمیک همکار هستند.



شکل ۱: دو ست از رأس‌ها در گراف نیت-ساینس نیومن (Newman's Net-science) نمایش داده شده است. ست سرخ ساختار یک جامعه را تجسم می‌کند، در حالی که ست آبی این‌گونه نیست (Abbe, 2018)

اگرچه شکل ۱ برای درک کلسترسازی گراف ما را به شکل شهودی کمک می‌کند؛ اما به طریق‌های مختلفی بیش از حد ساده‌سازی شده است. گراف‌های که در دنیای واقعی هستند معمولاً بسیار بزرگ‌تر اند و با یک طرح دو بعدی ساده که ساختار جامعه را به وضوح نشان می‌دهد، همراه نیستند. علاوه بر این، هرچند واضح است که ست سرخ در شکل ۱ بهتر از ست آبی به عنوان یک جامعه عمل می‌کند؛ اما در عمل فهمیدن این که چگونه بهترین تقسیم‌بندی یک گراف به جوامع را برای یک کاربرد خاص پیدا کنیم، بسیار چالش‌برانگیز است.

تعریف یک جامعه

کلسترسازی گراف به طور وسیع در رشته‌های مختلف مورد مطالعه قرار گرفته است و توافق عمومی درباره ویژگی‌های بنیادی یک جامعه گراف وجود دارد. در ذیل چندین نقل قول درباره تعریف یک جامعه را لیست می‌کنیم که از مقالات معتبر در این زمینه گرفته شده است. تمام این توضیحات می‌توانند برای درک بهتر اینکه چرا ست سرخ در شکل ۱ نمایانگر ساختار جامعه است، در حالی که ست آبی این‌گونه نیست، به کار روند.

- یک جامعه اغلب به عنوان ستای از رأس‌ها در نظر گرفته می‌شود که ارتباطات بیشتری بین اعضای آن نسبت به بقیه شبکه وجود دارد (Puleo & Milenkovic, 2015).
- کلسترسازی گراف وظیفه گروپ بندی رأس‌های گراف به کلسترها را بر عهده دارد و ساختار اضلاع گراف را به گونه‌ای در نظر می‌گیرد که باید تعداد زیادی اضلاع درون هر کلستر و نسبتاً تعداد کمی بین کلسترها وجود داشته باشد (Hou et al., 2016).
- یک ساختار که به آن جامعه گفته می‌شود، شامل گروپ از رأس‌ها است که به طور نسبی به یکدیگر متصل هستند اما به گروپ‌های متراکم دیگر در شبکه به صورت پراکنده متصل‌اند (Bhattacharya & De, 2008).
- جامعه‌ها یا کلسترها معمولاً گروپ‌هایی از رأس‌های هستند که احتمال بیشتری برای اتصال به یکدیگر نسبت به اعضای گروپ‌های دیگر دارند، هرچند الگوهای دیگری نیز ممکن است (Asteris et al., 2016).
- اساسی‌ترین وظیفه تشخیص جامعه یا کلسترسازی گراف شامل تقسیم رأس‌ها یک گراف به کلسترهایی است که به طور متراکم‌تری به هم متصل هستند (Yang & Leskovec, 2015).
- به طور کلی، هدف تخنیک‌های تقسیم‌بندی گراف و کلسترسازی گراف تشخیص ست‌های فرعی رأس‌ها با اضلاع داخلی زیاد و اضلاع خارجی کم است (Andersen et al., 2006).

این نقل‌قول‌ها دو راهنمای اساسی برای کلسترسازی گراف را برجسته می‌کنند:

1. جوامع باید کثافت داخلی بالایی داشته باشند، یعنی تعداد زیادی اضلاع بین رأس‌های درون یک جامعه وجود داشته باشد.
2. جوامع باید اتصالات خارجی پراکنده‌ای داشته باشند، یعنی نباید تعداد زیادی اضلاع که یک جامعه را به بقیه گراف متصل کند وجود داشته باشد.

تقریباً هر رویکردی در کلسترسازی گراف به نوعی بر این دو هدف (که اغلب متضاد هستند) تکیه دارد. برای یافتن کلسترهایی با کثافت داخلی بالاتر، ممکن است لازم شود رأس‌هایی را که اضلاع زیادی دارند اما شاید برای یک جامعه کافی نباشند، را حذف کنیم. هرچه ما از یک جامعه بخواهیم متراکم‌تر باشد، بیشتر مجبور به حذف اینگونه رأس‌ها خواهیم شد. این به نوبت خود منجر به افزایش تعداد اضلاع بین جوامع می‌شود. از سوی دیگر، اگر بخواهیم به ساده گی از اضلاع بین جوامع اجتناب کنیم، می‌توانیم همه رأس‌ها را در یک کلستر قرار دهیم. اما، این از نظر ریاضی بی‌اهمیت و در کاربردها

غیرمفید خواهد بود. بنابراین، درک چگونگی ایجاد تعادل صحیح بین کثافت داخلی و پراکندگی خارجی یکی از اساسی‌ترین تصمیمات در کلسترسازی گراف است.

کلسترسازی گراف به عنوان بهینه‌سازی ترکیبی

ما به کلسترسازی گراف از طریق استراتژی مشترک بهینه‌سازی ترکیبی نزدیک می‌شویم. به عبارت دیگر، این مسأله با معرفی یک تابع هدف که یک امتیاز عددی به هر کلستر گسسته از یک گراف اختصاص می‌دهد، فورمول بندی می‌شود و اندازه‌گیری می‌کند که چقدر خوب ساختار یک جامعه را تجسم می‌کند. بهینه‌سازی تابع هدف بر روی تمام کلسترهای ممکن، بهترین تقسیم‌بندی گراف را نسبت به یک معیار مشخص بازمی‌گرداند. به عنوان مثال، یکی از راه‌های تقسیم‌بندی یک گراف G به دو بخش، تشخیص ست‌ای از رأس‌های S است که تابع f را مینیمم می سازد، که به آن تابع هدف قطع نرمال شده گفته می‌شود (Yu & Ding, 2010):

$$f(S) = \frac{\text{تعداد اضلاع خارج شده از S}}{\text{تعداد اضلاع سرحدی که در S نیستند}} + \frac{\text{تعداد اضلاع خارج شده از S}}{\text{تعداد اضلاع سرحدی در S}} \qquad (1)$$

مینیمم کردن f دو کلستر (رأس‌های در S و رأس‌های خارج از S) را تولید می‌کند که هر دو در اندازه غیرقابل مقایسه هستند و تعداد کمی ضلع با یکدیگر دارند. رابطه (۱) طور فورمول بندی شده است که کلسترها را به سوی کثافت داخلی بالاتر اضلاع سوقت می‌دهد. با شامل ساختن "تعداد اضلاع خارج شده از S" در صورت‌های کسری تضمین می‌کند که خروجی بهینه تعداد کمی ضلع با بقیه گراف دارد، یعنی پراکندگی خارجی تشویق می‌شود. با نگاهی به شکل ۱، اگر S_r را ست رأس‌های سرخ و S_b را به عنوان رأس‌های آبی فرض کنیم ، یک محاسبه سریع نشان می‌دهد که $f(S_r) \approx 0.0119$ و $f(S_b) \approx 1.0319$. واقعیت این که امتیاز قطع نرمال شده S_r بسیار کمتر از امتیاز S_b است، با شهود ما مطابقت دارد که S_r یک کلستر خوب بوده و S_b نیست.

استراتژی‌های مرتبط برای کلسترسازی گراف

بهینه‌سازی یک تابع هدف تنها استراتژی موجود برای کلسترسازی گراف نیست. فورتوناتو و هریک (Fortuato & Hric, 2016) نمای کلی از روش‌های بهینه‌سازی را در نظرسنجی خود درباره کلسترسازی گراف ارائه دادند. علاوه بر این، آن‌ها نمای کلی از روش‌های احصاییه استنباطی و روش‌های دینامیکی را نیز ارائه دادند. به طور خلاصه، روش‌های احصاییه استنباطی وجود یک توزیع زیرین از گراف‌ها را با ساختار جامعه فرض می‌کنند. کلسترسازی یک گراف معادل است با استنباط پارامترهای توزیع و یادگیری تخصیص کلسترای که احتمال مشاهده گراف داده شده را به حداکثر

می‌رساند. روش‌های دینامیکی برای کلسترسازی گراف بر این مفکوره تکیه دارند که یک پروسه دینامیکی در شبکه، مانند یک پیاده‌روی تصادفی یا نوع دیگری از انتشار، زمانی که با یک کلستر متصل به هم مواجه می‌شود، در محل خود باقی خواهد ماند. به طور واضح‌تر، اگر یک پیاده‌رو تصادفی در یک گراف با رأس‌های که توسط اضلاع به شکل اتفاقی تعقیب می‌شوند، مقابل شود ، به احتمال زیاد در داخل آن جوامع گرفتار خواهد شد. ما انتظار داریم که پیاده‌رو برای مدت طولانی در اطراف بسیار از اتصالات داخلی یک کلستر گردش نماید، قبل از این که یکی از (چند) ضلع خارجی را دنبال کرده و از کلستر فرار کند. زمانی که یک پروسه دینامیکی به این صورت گرفتار می‌شود، مشاهده یکی از راه‌های تشخیص جوامع است.

هر یک از این رویکردها، شیوه‌ای متفاوت از تفکر درباره کلسترسازی گراف را تشکیل می‌دهد. با این حال، آن‌ها به هیچ وجه متناقض یا متعارض نیستند، بلکه در واقع از نظر بنیادی به یکدیگر نزدیک هستند. پیاده‌روی‌های تصادفی در گراف‌ها به طور نزدیکی به آرامش‌های طیفی توابع هدف که ممکن است بخواهید بهینه‌سازی کنید، مرتبط هستند (Lu, 2002). علاوه بر این، بسیاری از روش‌های مبتنی بر انتشار کلسترهایی را تولید می‌کنند که تضمین‌های قوی نسبت به اهدافی مانند هدایت دارند (Wang et al., 2017). همچنین معادلات شناخته شده‌ای بین تخصیص جوامع بر اساس مودل بلوک تصادفی (Abbe, 2018). (روشی مبتنی بر احصاییه استنباطی) و اهداف رایج کلسترسازی گراف مانند قطع‌کننده حداقل و مکزیمم مدولاریتی وجود دارد (Neewman, 2016). در نهایت، نشان داده شده است که کلسترسازی با استفاده از Personalized PageRank که یک روش دینامیکی است ارتباطات قوی با استنباط ساختار جامعه بر اساس مودل بلوک تصادفی دارد (Kloumann et al., 2017). این نتایج تنها نمونه‌ای از ارتباطات عمیق موجود بین روش‌های مبتنی بر بهینه‌سازی و سایر رویکردهای موجود هستند.

الگوریتم‌ها و پیچیدگی محاسباتی

تعریف ساختار جامعه تنها اولین قدم در تشخیص واقعی جوامع است. با توجه به یک تابع هدف، قدم حیاتی بعدی توسعه تخنیک‌های معقول برای بهینه‌سازی آن است. بنابراین توسعه الگوریتم و مطالعه پیچیدگی محاسباتی اجزای کلیدی در تحقیق کلسترسازی گراف هستند.

بیشتر توابع هدف برای کلسترسازی گراف که معرفی و تحلیل شده‌اند، به عنوان NP-hard شناخته می‌شوند، به این معنی که بهینه‌سازی دقیق آن‌ها در عمل غیرممکن است. تمرکز اصلی در کلسترسازی گراف توسعه الگوریتم‌های کارآمد برای تشخیص جوامع است که به سرعت اجرا می‌شوند و می‌توانند

ساختار کلسترسازی معناداری را در تیوری و عمل تشخیص کنند. بسیاری از این الگوریتم‌ها برای بهینه‌سازی یک تابع خاص طراحی شده‌اند، که معمولاً با انجام حرکات محلی به منظور بهبود امتیاز تابع هدف کلسترسازی می‌کنند. با این حال، این تخنیک‌ها معمولاً هیچ تضمینی برای کیفیت تقریبی تابع هدف ارائه نمی‌دهند و ممکن است در بدترین حالت عملکرد ضعیفی داشته باشند. رویکردی دقیق‌تر برای کلسترسازی گراف توسعه الگوریتم‌های تقریبی کارآمد برای اهداف NP-hard است. این الگوریتم‌ها در زمان پولینومیی اجرا می‌شوند و کلستر ای را تولید می‌کنند که به طور قابل اثبات نزدیک به راه‌حل بهینه در یک عامل ضربی خاص است. الگوریتم‌های تقریبی در زمینه علوم کامپیوتر به طور وسیع‌ای مطالعه می‌شوند و الگوریتم‌های شناخته‌شده‌ای برای بسیاری از اهداف کلسترسازی گراف وجود دارد.

درک پیچیدگی محاسباتی یک مسأله کلسترسازی گراف خاص همچنین شامل اثبات نتایج نظری درباره محدودیت‌های بنیادی مرتبط با حل یا تقریب یک تابع هدف است. واقعیت این است که بیشتر اهداف کلسترسازی NP-hard هستند و نشان می‌دهد که تنها الگوریتم‌های شناخته شده برای حل بهینه این اهداف در زمان نمایی اجرا می‌شوند. همچنین مفاهیم دقیق‌تری از پیچیدگی محاسباتی وجود دارد که می‌توان از آن‌ها برای نشان دادن اینکه حتی به دست آوردن انواع خاصی از تقریب‌ها چالش‌برانگیز است، استفاده کرد. جزئیات بیشتری را در فصل بعد ارائه خواهیم داد.

گراف‌ها و قطع‌کننده‌ها

در این مقاله، ما از $G = (V, E)$ برای نشان دادن یک گراف بدون وزن و بدون جهت طوریکه $n = |V|$ رأس‌ها و $m = |E|$ اضلاع را نشان می‌دهد، استفاده خواهیم کرد. همچنان فرض می‌کنیم که گراف‌ها حلقه ندارند و با چند استثنا تنها گراف‌های متصل را در نظر می‌گیریم. درجه یک رأس $v \in V$ تعداد اضلاع حادث (وابسته) بر روی v است که به d_v نشان داده می‌شود. برای دو ست مجزا از رأس‌های $S, T \subseteq V$، $cut(S,T)$ نشان دهنده تعداد اضلاع بین آن‌ها است. برای هر ست از رأس‌های $S \subseteq V$، ما $\bar{S} = V \backslash S$ را به عنوان ست مکمله آن تعریف می‌کنیم و $E_S \subset E$ نشان دهنده ست اضلاع در گراف فرعی القایی S است. همچنین رابطه‌های ذیل نیز صحت است:

$$\begin{aligned} \text{cut}(S) &= \text{cut}(S, \bar{S}) \\ \text{vol}(S) &= \sum_{v \in S} d_v \\ \text{density}(S) &= \frac{|E_S|}{\binom{|S|}{2}} \end{aligned} \tag{2}$$

برای یک رأس منفرد $v \in V$ ، $\text{density}(v) = 1$ را در نظر می‌گیریم. از آنجا که هر ست $S \subset V$ به طور منحصر به فرد با ست‌ای از اضلاع بین خود و بقیه گراف تشخیص می‌شود، ما اغلب به یک ست از رأس‌ها به عنوان یک قطع‌کننده (cut) در یک گراف اشاره خواهیم کرد.

کلسترسازی گراف

ما از $\mathcal{C}$ برای نشان دادن یک کلستر در گراف استفاده می‌کنیم. برای ما در این مقاله، کلسترسازی با تقسیم رأس‌ها معادل است، یعنی $\mathcal{C} = \{S_1, S_2, \dots, S_k\}$ یک $k - clustering$ است اگر $S_i \cap S_j = \emptyset$ برای هر $i \neq j$ و $\bigcup_{i=1}^{k} S_i = V$ باشد. ما یک کلستر $\mathcal{C}$ از یک شبکه را با استفاده از یک تابع شاخص کلسترسازی $\delta^{\mathcal{C}} = \left(\delta_{ij}^{\mathcal{C}}\right)$ کدگذاری می‌کنیم.

$$\delta_{ij}^{\mathcal{C}} = \begin{cases} 1 & \text{اگر رأس‌های } i \text{ و } j \text{ در یک کلستر باشند} \\ 0 & \text{اگر رأس‌های } i \text{ و } j \text{ در یک کلستر نباشند} \end{cases} \quad (3)$$

ما همچنین کلسترها را با استفاده از ست متحول‌ها $\mathbf{x}^{\mathcal{C}} = \left(x_{ij}^{\mathcal{C}}\right)$ کدگذاری می‌کنیم که در آن $x_{ij}^{\mathcal{C}} = 1 - \delta_{ij}^{\mathcal{C}}$ نشان دهنده فاصله باینری بین دو رأس i و j است (i و j فاصله ۱ دارند اگر جدا از هم کلسترسازی شده باشند و در غیر این صورت فاصله ۰ دارند). در این مقاله، کلسترسازی گراف برای حل بهینه‌سازی یک تابع هدف $f: \mathfrak{C} \to \mathbb{R}$ مطرح می‌شود، جایی که $\mathfrak{C}$ نشان دهنده ست‌ای از کلسترهای درست (معتبر) است. برای هر کلستر $\mathcal{C} \in \mathfrak{C}$ ، تابع f یک امتیاز از تابع هدف یعنی $f(\mathcal{C}) \in \mathbb{R}$ را خروجی می‌دهد که معیاری از این که چقدر خوب $\mathcal{C}$ ساختار جامعه را تجسم می‌کند، ارائه می‌دهد.

کلسترسازی، تقسیم‌بندی و تشخیص جامعه

اصطلاحات کلسترسازی گراف، تشخیص جامعه و تقسیم‌بندی گراف همه در بخش مفاهیم اساسی به کار رفتند. کلمه‌های کلستر و جامعه معمولاً به طور متناوب برای اشاره به رأس‌های به‌هم‌پیوسته‌ای که تنها به‌طور ضعیف به بقیه گراف متصل هستند، استفاده می‌شوند. بر اساس استانداردهای موجود در ادبیات، ما در این مقاله از اصطلاحات تشخیص جامعه و کلسترسازی گراف به شکل مترادف استفاده خواهیم کرد. هر چند تقسیم‌بندی گراف نیز مشابه به کلسترسازی و تشخیص جامعه است؛ اما در چندین جنبه کلیدی از این اصطلاحات متمایز می‌شود. ما این تفاوت‌ها را قبل از تعاریف رسمی‌تر برجسته می‌کنیم.

تقسیم‌بندی گراف معمولاً در زمینه محاسبات موازی ارائه می‌شود (Christian, 2013)، جایی که رأس‌ها در یک گراف نشان دهنده وظایف محاسباتی هستند و اضلاع نوعی وابستگی دیتاها را نشان

می‌دهند. در این حالت، اگر k پروسسر برای انجام کار وجود داشته باشند، مفید است که وظایف را به طور دقیق به k کلستر تقسیم کنیم به گونه‌ای که ارتباط بین پروسسرها حداقل شود. برای اینکه کار بین k پروسسر متعادل باشد، مهم است که رأس‌ها را به بلوک‌هایی با اندازه تقریباً یکسان تقسیم کنیم. مطابق با این مثال، بیشتر تقسیم‌بندی گراف تعداد k کلستر را برای تشکیل مشخص می‌کند و یک قید تعادل بر روی اندازه‌های کلستر را شامل می سازد (Schulz et al., 2016)، (Fortuato & Hric, 2016). پس هدف اینجا یافتن k طریقه برای کلسترسازی است که قید تعادل را برآورده کند و تعداد اضلاع بین کلسترها را حداقل کند. با توجه به تأکید بر حداقل کردن ارتباطات، تقسیم‌بندی گراف معمولاً اولویت بیشتری بر پراکندگی خارجی نسبت به کثافت داخلی می‌دهد.

اهداف طراحی شده تقسیم‌بندی گراف گاهی با کاربردهای کلسترسازی گراف و تشخیص جامعه در تضاد است. در بسیاری از شبکه‌های دنیای واقعی (مانند شبکه‌های بیولوژیکی یا اجتماعی)، فرض اینکه تعداد ثابتی از کلسترها وجود دارد، مفید نیست. علاوه بر این، ممکن است جوامع طبیعی در گراف از نظر اندازه متفاوت باشند. به همین دلایل، تقسیم‌بندی گراف معمولاً با کاربردهایی در محاسبات با کارایی بالا (Andersen et al., 2006)، ضرب متریکس‌های پراکنده (Asteris et al., 2016) و طراحی $VSLI$ (Bader et al., 2013) مرتبط است. در بسیاری از این کاربردها، گراف دارای ساختار هندسی بسیار خاص است که با شبکه‌های پیچیده و چندبعدی که معمولاً هدف مطالعه در کلسترسازی گراف هستند، متفاوت است.

تفاوت‌های ظریف بین کلسترسازی گراف و تقسیم‌بندی گراف در تعدادی از سروی‌ها (Chierichetti et al., 2014) و تیزس‌ها (Christian, 2013) به تفصیل مورد بحث قرار گرفته است. با وجود این تفاوت‌های موجود، نقاط مشترک بین تخنیک‌های تقسیم‌بندی گراف و ابزارهای تشخیص جامعه وجود دارد و مسائل اغلب در ادبیات به‌طور مشترک ارائه و مطالعه می‌شوند (Bader et al., 2013)، (Newman, 2013) (Wang et al., 2015)، (Zhou et al., 2017)، (Schulz et al., 2016). اگرچه ما عمدتاً بر کلسترسازی گراف و تشخیص جامعه تمرکز داریم؛ اما بسیاری از نتایج به‌طور مستقیم به تقسیم‌بندی گراف نیز قابل اعمال خواهند بود.

اصطلاحات پیچیدگی محاسباتی

با توجه به تأکید ما بر کلسترسازی گراف مبتنی بر بهینه‌سازی، به‌طور خلاصه اصطلاحات استاندارد در مورد الگوریتم‌های تقریبی و نتایج برای مسائل محاسباتی را مرور می‌کنیم. یک مسأله تصمیم‌گیری، یک مسأله محاسباتی است که تنها به یک پاسخ بلی یا نخیر نیاز دارد. بخش تصمیم‌گیری برای هر

مسأله کلسترسازی گراف مبتنی به بهینه‌سازی می‌پرسد که آیا برای یک امتیاز هدف ثابت β ، کدام کلستر C وجود دارد که امتیاز هدف آن کمتر از β باشد؟ یک مسأله تصمیم گیری در P است اگر بتوان آن را در زمان چندجمله‌ای حل کرد. یک مسأله تصمیم گیری در NP است اگر بتوان آن را در زمان چندجمله‌ای بررسی کرد، یعنی با توجه به یک کلستر C و یک امتیاز β ، آیا کلستر C امتیاز کمتری از β دارد؟ یک مسأله $NP-hard$ است اگر هر مسأله‌ای در NP بتواند در زمان چندجمله‌ای به آن کاهش یابد. مسائل $NP-complete$ آن‌هایی هستند که هم در $NP-hard$ و هم در NP قرار دارند.

چندین مفهوم دیگر از پیچیدگی وجود دارد که مفید است به آن‌ها اشاره کنیم. ما این‌ها را به‌طور خاص برای مسائل مینیمم‌سازی تعریف خواهیم کرد، هرچند با تغییرات ساده‌ای می‌توان برای انواع مکزیمم‌سازی نیز حل کرد. یک الگوریتم با ضریب ثابت برای یک مسأله، الگوریتمی است که در زمان چندجمله‌یی اجرا می‌شود و راه‌حلی را در محدوده‌ی یک ضریب ثابت از بهینه ارائه می‌دهد، یعنی برای یک مسأله مینیمم‌سازی، یک الگوریتم تقریبی c راه‌حلی را به ما می‌دهد که حداکثر $c \cdot OPT$ باشد که در آن OPT امتیاز بهینه است. یک الگوریتم، طرح تقریب زمان چندجمله‌ای $(PTAS)$ نامیده می‌شود اگر برای هر $\varepsilon > 0$ ثابت، الگوریتمی‌در زمان چندجمله‌ای وجود داشته باشد (که زمان اجرای آن معمولاً وابسته به ε است) که یک تقریب $(1+\varepsilon)$ را بازمی‌گرداند. یک مسأله $APX-hard$ گفته می‌شود اگر یک ثابت $c > 1$ وجود داشته باشد به طوری که $NP-hard$ آن را در یک ضریب کوچکتر از c تقریب بزند. بنابراین $PTAS$ وجود ندارد مگر اینکه $P = NP$. فرضیه بازی‌های منحصر به فرد یک مشکل چالش‌برانگیز و باز در علم کامپیوتر است که توسط خوت مطرح شده و مربوط به $NP-complete$ از مسائل بازی‌های منحصر به فرد است. یک مسأله $UG-hard$ گفته می‌شود اگر $NP-hard$ باشد، با فرض اینکه فرضیه بازی‌های منحصر به فرد صحیح باشد.

توابع هدف کلسترسازی گراف

توابع هدف متعددی در بخش کلسترسازی گراف معرفی و به‌طور کامل تحلیل شده‌اند. در اینجا سه دسته از توابع هدف را با مثال‌های آن در نظر می‌گیریم، که هر یک تعادل متفاوتی بین کثافت داخلی و پراکندگی خارجی کلسترهای تشکیل‌شده برقرار می‌کند.

اهداف نسبت قطع‌کننده به اندازه

بسیاری از توابع هدف به‌طور خاص برای اندازه‌گیری ساختار اجتماعی یک ست $S \subset V$ به اساس نسبت بین $\text{cut}(S)$ و اندازه ست تعریف شده‌اند. پراکندگی یک قطع‌کننده $S \subset V$ به‌صورت زیر تعریف می‌شود:

$$\text{scut}(S) = \frac{\text{cut}(S)}{|S|} + \frac{\text{cut}(\bar{S})}{|\bar{S}|} = n \cdot \frac{\text{cut}(S)}{|S||\bar{S}|} \quad (4)$$

یافتن ست S با حداقل پراکندگی به عنوان مساله قطع‌کننده کم‌پراکندگی شناخته می‌شود و این یک مساله محبوب در علم کامپیوتر است. نه تنها این یک مساله NP-hard است، بلکه به‌طور خاص UG-hard نیز است که بتوان آن را در هر عامل ثابتی تقریب زد (Chawla et al., 2015). برای سال‌ها، بهترین نسبت تقریب شناخته‌شده $O(\log n)$ بود که توسط Leighton و Rao ارائه شد. بعداً این نسبت به $O(\sqrt{\log n})$ توسط (Arora et al., 2009) با گرد کردن یک برنامه نیمه‌تعریف‌شده بهبود یافت. اغلب راحت‌تر است که با نسخه مقیاس‌بندی شده‌ای از هدف کار کنیم که تعداد کل اضلاع بین S و $\bar{S}$ (یعنی $\text{cut}(S)$) را اندازه‌گیری کرده و بر حداکثر تعداد ممکن اضلاع بین دو ست (یعنی $|S||\bar{S}|$) تقسیم کنیم. ما به این مقدار به عنوان پراکندگی مقیاس‌بندی شده S اشاره می‌کنیم که آن را به صورت زیر نشان می‌دهیم:

$$\text{sscut}(S) = \frac{\text{cut}(S)}{|S||\bar{S}|} = \frac{1}{n}\text{scut}(S) \quad (5)$$

تعدادی دیگر از اهداف به‌طور نزدیک با قطع‌کننده کم‌پراکندگی مرتبط هستند. گسترش یک ست به‌صورت زیر تعریف می‌شود:

$$\text{expan}(S) = \frac{\text{cut}(S)}{\min\{|S|, |\bar{S}|\}} \quad (6)$$

که از منظر پیچیدگی محاسباتی به‌طور مکرر مطالعه می‌شود. نسخه‌های دارای وزن و درجه این مسایل، قطع‌کننده نورمال‌شده و امتیاز هدایت شده هستند که به‌ترتیب به صورت زیر تعریف می‌شوند:

$$\text{ncut}(S) = \frac{\text{cut}(S)}{\text{vol}(S)} + \frac{\text{cut}(\bar{S})}{\text{vol}(\bar{S})} \quad (7)$$

$$\text{cond}(S) = \frac{\text{cut}(S)}{\min\{\text{vol}(S), \text{vol}(\bar{S})\}} \quad (8)$$

قطع‌کننده نرمال‌شده توسط شولز در زمینه قطعه‌بندی تصویر معرفی شد (Christian, 2013)، در حالی که هدایت به عنوان یکی از رایج‌ترین اهداف برای تشخیص جامعه شناخته می‌شود (Kloumann et

(al., 2017. از منظر نظری، همه این اهداف تقریباً یکسان در نظر گرفته می‌شوند. برای گراف‌های درجه دار و منظم، هدایت و گسترش تا یک عامل مضرب ثابت یکسان هستند، همانطور که قطع‌کننده کم‌پراکندگی و قطع‌کننده نرمال‌شده هستند.

رابطه با اهداف چندکلستری

اگر چه اهداف که در فوق ذکر شد به طور خاص برای یک ست واحد S تعریف شده اند؛ اما هنوز هم به عنوان اهداف کلسترسازی جهانی در نظر گرفته می‌شوند، زیرا تشخیص یک جامعه واحد S معادل تشکیل یک کلستر دوگانه از شبکه است: $\mathcal{C} = \{S, \bar{S}\}$. تعمیم‌های چند کلستر ای از این اهداف نیز مورد بررسی قرار گرفته است. به‌عنوان مثال، زمانی که شولز امتیاز قطع‌کننده نرمال‌شده را معرفی کرد، نسخه‌یی $k - way$ از هدف را نیز ارائه داد (Christian, 2013).

$$\mathrm{ncut}_k(\mathcal{C}) = \frac{\mathrm{cut}(S_1, \bar{S}_1)}{\mathrm{vol}(S_1)} + \frac{\mathrm{cut}(S_2, \bar{S}_2)}{\mathrm{vol}(S_2)} + \cdots + \frac{\mathrm{cut}(S_k, \bar{S}_k)}{\mathrm{vol}(S_k)} \qquad (9)$$

برای یک k کلستر $\mathcal{C} = \{S_1, S_2, \cdots, S_k\}$

اهداف مبتنی بر مودولاریتی

به‌طور قابل‌توجهی، رایج‌ترین هدف برای تشخیص جامعه، نمره مودولاریتی نیومن و گیرون است (Newman, 2013). به‌طور شهودی، گفته می‌شود که یک کلستر دارای مودولاریتی بالا است اگر کلسترها کثافت داخلی اضلاع بالاتری نسبت به آنچه که به‌طور تصادفی انتظار می‌رود داشته باشند. تصادفی بودن در این زمینه توسط یک مودل خنثی از پیش مشخص‌شده تعریف می‌شود که احتمال P_{ij} وجود یک ضلع بین رأس‌های i و j را تعیین می‌کند. به‌طور رسمی، مودولاریتی یک کلستر C به‌صورت زیر تعریف می‌شود:

$$\mathrm{mod}(\mathcal{C}) = \frac{1}{2m}\sum_{i=1}^{n}\sum_{j=1}^{n}\left(A_{ij} - P_{ij}\right)\delta_{ij}^{\mathcal{C}} \qquad (10)$$

تعدادی از انتخاب‌های مختلف برای مودل خنثی مودولاریتی مورد بررسی قرار گرفته است. ابتدایی‌ترین آن مودل گراف تصادفی برنولی است که در آن انتظار وجود یک ضلع برابر با $P_{ij} = p$ برای بعضی از $p \in (0,1)$ و برای همه جوره‌های رأس‌های $i \neq j$ است. با این حال، رویکرد بسیار محبوب‌تر استفاده از مودل خنثی فن-چونگ-لو (Fan, 2009) است که با تنظیم $P_{ij} = d_i d_j/(2m)$ تعریف می‌شود. انتخاب این مودل تضمین می‌کند که تعداد اضلاع مورد انتظار برابر با تعداد اضلاع واقعی در گراف مشاهده شده G است، زیرا $\sum_{i\neq j} P_{ij} = \sum_{i\neq j} A_{ij}$. علاوه بر این،

توزیع درجه حفظ می‌شود. به عبارت دیگر، می‌توان نشان داد که $\sum_{j\neq i} P_{ij} = d_i$ برای هر رأس i، بنابراین درجه مورد انتظار رأس i در گراف تصادفی برابر با درجه رأس i در G است.

محدودیت‌های مودولاریتی

مودولاریتی به‌طور وسیع‌ای مورد استفاده قرار گرفته و انواع مختلفی هدف برای گراف‌های وزنی، گراف‌های دو قسمتی، شبکه‌های چندلایه و چندین مورد دیگر معرفی شده است. با این حال، با وجود استفاده وسیع، دارای چندین محدودیت نیز است. اول از همه، حتی گراف‌های تصادفی ممکن است نمرات مودولاریتی بالایی را نشان دهند که تشخیص اینکه آیا نمره مودولاریتی بالا نشان‌دهنده ساختار معنادار است یا خیر را دشوار می‌کند (Kloumann et al., 2017). همچنین مشخص شده است که مودولاریتی از یک محدودیت ذاتی در دقت رنج می‌برد (Fortuato & Hric, 2016)، به این معنی که ممکن است نتواند جوامعی را که کوچکتر از اندازه خاصی هستند تشخیص کند، که به اندازه شبکه بستگی دارد. در نهایت، بهینه‌سازی مودولاریتی بسیار چالش برانگیز است. نه تنها هدف $NP-hard$ است، بلکه دین و همکاران (Dau et al., 2017) نشان دادند که حتی تقریب‌زدن آن به یک عامل ثابت نیز $NP-hard$ است. بنابراین، اگرچه روش‌های بسیاری به صورت ابتکاری معرفی شده‌اند، هیچ یک از آن‌ها تضمین‌های قابل اثباتی برای تقریب ندارند.

مودولاریتی‌های تعمیم یافته

رایشاردت و بورنهولدت تعمیمی از مودولاریتی را بر اساس یافتن حالت‌های مینیمم انرژی مودل شیشه اسپین معرفی کردند (Pan et al., 2015). برای اهداف ما کافی است که درک کنیم که رویکرد آن‌ها به کلسترسازی گراف معادل با منیمم کردن تابع هدف‌هامیلتونی زیر است:

$$\text{Hamiltonian}(\mathcal{C}) = -\sum_{i\neq j} \left(A_{ij} - \gamma P_{ij}\right)\delta_{ij}^{\mathcal{C}} \qquad (11)$$

که در آن γ یک پارامتر دقت قابل تغییر در کلسترسازی است. زمانی که $\gamma = 1$، مینیمم کردن‌هامیلتونی معادل با ماکزیمم کردن مودولاریتی است.

دلوین و همکاران تعمیم دیگری از مودولاریتی را تحت عنوان ثبات یک کلستر معرفی کردند (Delvenne et al., 2010). ثبات یک قسمت احتمال این‌که یک راهرو تصادفی در یک گراف چی وقت به یک راهپیمایی تصادفی به طول t در کلستر ای که در آن شروع کرده است پایان می‌دهد را اندازه گیری می‌کند. طول راهپیمایی t نوع دیگری از پارامتر دقت قابل تغییر است که می‌تواند برای تشخیص انواع مختلف کلستر در یک گراف تغییر کند. با افزایش t، راهرو تصادفی "دورتر" از نقطه شروع اولیه خود خواهد رفت و هدف تمایل دارد تا تشخیص کلسترهای بزرگ‌تر در گراف را پاداش

دهد. دلوین و همکاران اثبات کردند که نسخه خطی شده ای از ثبات معادل با تابع هدف‌هامیلتونی (۲,۹) رایشاردت و بورنهولدت است (Delvenne et al., 2010).

هر دو هدف ثبات و هدف‌هامیلتونی (۲,۹) مودولاریتی را تعمیم می‌دهند و راهی برای غلبه بر محدودیت دقت با اجازه دادن به کاربران برای تنظیم یک پارامتر دقت (γ یا t) فراهم می‌کنند. با این حال، بهینه‌سازی این اهداف همچنان بسیار چالش برانگیز باقی مانده و تمام الگوریتم‌های شناخته شده تا به حال که ابتکاری هستند هیچ تضمینی برای تقریب ندارند.

اهداف تغییریافته گراف کلستر شده

تعریف ایده‌آل از یک جامعه در یک گراف، یک ستی از رأس‌های متصل است که هیچ ضلع مشترک با بقیه گراف ندارد. یک گراف $G = (V, E)$ که به طور کامل از کلایک‌های مجزا تشکیل شده باشد، به عنوان یک گراف کلستر ای شناخته می‌شود. بر اساس این مفهوم ایده‌آل از ساختار جامعه، اهداف تغییریافته‌ی گراف کلستر شده تعداد اضلاع‌هایی را که باید در یک گراف تغییر یابند تا آن را به یک گراف کلستر ایم تبدیل کنند، اندازه‌گیری می‌کنند. یک نوع از این مسأله نخستین بار توسط بن-دور و همکاران در زمینه کلسترسازی الگوهای بیان ژن مورد مطالعه قرار گرفت (Ben-Dor et al., 1999). شامیر و همکاران بعداً سه هدف مرتبط را به صورت رسمی معرفی کردند: تکمیل کلستر، ویرایش کلستر و حذف کلستر (Sharma & Singh, 2016).

تعاریف رسمی

تکمیل کلستر به دنبال حداقل تعداد اضلاع‌هایی است که باید به یک گراف اضافه شود تا آن را به یک گراف کلستری تبدیل کند. این مسأله می‌تواند در زمان چندجمله‌یی حل شود؛ زیرا معادل پیدا کردن مؤلفه‌های متصل در یک گراف است. حذف کلستر به دنبال حداقل تعداد اضلاع‌هایی است که باید از یک گراف حذف شود تا آن را به یک اتحاد مجزا از کلایک‌ها تبدیل کند. این معادل با پیدا کردن یک کلستر $\mathcal{C}$ است که در آن تمام کلسترها کلایک هستند و تعداد اضلاع‌های بین این کلایک‌ها حداقل است. به‌طور رسمی، هدف به صورت زیر نوشته می‌شود:

$$\begin{aligned} \min_{\mathcal{C}} \quad & \sum_{i<j} A_{ij}\left(1-\delta_{ij}^{\mathcal{C}}\right) \\ \text{subject to} \quad & \delta_{ij}^{\mathcal{C}} = 0 \text{ if } (i,j) \notin E \end{aligned} \tag{12}$$

نتانزون نشان داد که حذف کلستر برای بهینه‌سازی NP-hard است (Ben-Dor et al., 1999)، در حالی که شامیر و همکاران ثابت کردند که در واقع APX-hard است. ویرایش کلستر اجازه اضافه و

حذف اضلاع را می‌دهد و توسط شامیر و همکاران نشان داده شد که NP-complete است. به طور رسمی، هدف به صورت زیر است:

$$\min_{\mathcal{C}} \sum_{i<j} A_{ij}\left(1-\delta_{ij}^{\mathcal{C}}\right)+\left(1-A_{ij}\right)\delta_{ij}^{\mathcal{C}} \qquad (13)$$

برای یک کلستر ثابت $\mathcal{C}$ ، از $\mathrm{cedit}(\mathcal{C})$ در ویرایش کلستر و $\mathrm{cdel}(\mathcal{C})$ **در** حذف کلستر برای نشان دادن تعداد اضلاع‌هایی که باید در گراف تغییر یابند، استفاده می‌کنیم، تا $\mathcal{C}$ به یک ست مجزای از کلایک‌ها تبدیل شود. در حذف کلستر، برای هر کلستر در $\mathcal{C}$ که یک کلایک نباشد، نمره $\mathrm{cdel}(\mathcal{C})=\infty$ را اختصاص می‌دهیم.

نتایج پیچیدگی

تعدادی از نتایج پیچیدگی پارامتری برای ویرایش و حذف کلستر ارائه شده است (Bocker & Baumbach, 2013). الگوریتم‌های پارامتری یک بودجه ثابت k را در نظر می‌گیرند و به دنبال پاسخ بلی یا نخیر برای اینکه آیا می‌توان یک گراف را با استفاده از حداکثر k تغییرات به یک گراف کلسترای تبدیل کرد، هستند. بوکر و بمباچ مروری بر نتایج قابلیت دسترسی پارامتر ثابت به‌طور خاص برای ویرایش کلستر ارائه داده‌اند. بسیاری از الگوریتم‌های زمان چند جمله ای و نتایج سختی برای صنف‌های خاصی از گراف‌ها نیز ارائه شده است (Boonomo et al., 2015a).

ویرایش کلستر معادل با کلسترسازی همبستگی بدون وزن است که ما در بخش بعدی به تفصیل درباره آن بحث خواهیم کرد. بنابراین بسیاری از نتایج کلسترسازی همبستگی مستقیماً به ویرایش کلستر منتقل می‌شوند. به طور خاص ثابت شده که این مسأله برای بهینه‌سازی APX-hard است (Charikar et al., 2017) و بهترین عامل تقریب کمی بهتر از ۲,۰۶ است (Chawla et al., 2015). حذف کلستر می تواند به عنوان نسخه ای محدود شده از کلسترسازی همبستگی دیده شود. چاریکار و همکاران نشان دادند که یک الگوریتم تقریب ۴ برای کلسترسازی همبستگی بدون وزن (یعنی ویرایش کلستر) می‌تواند برای به دست آوردن یک تقریب ۴ برای حذف کلستر سازگار شود. بعداً، وانزولین و ویلیمسون یک تقریب ۳ برای نسخه‌های محدود شده از کلستربندی همبستگی ثابت کردند که مستقیماً یک تقریب ۳ برای حذف کلستر را نتیجه می‌دهد.

پیش زمینه کلسترسازی همبستگی

کلسترسازی همبستگی چارچوبی برای کلسترسازی ست‌های از دیتاها است که با روابط جوره‌یی مثبت و منفی بین اشیاء دیتا مشخص می‌شود. این مسأله معمولاً به عنوان یک مسأله تقسیم‌بندی در گراف‌های علامت دار ظاهر می‌شود. در این زمینه، یک ضلع منفی در گراف نشان‌دهنده‌ی شواهدی است که دو

رأس باید از هم جدا شوند و یک ضلع مثبت نشان‌دهنده‌ی این است که دو رأس باید با هم کلستر شوند. از آن‌جا که قطعات فردی شواهد ممکن است با یکدیگر تضاد داشته باشند، هدف یافتن یک کلستر از دیتاها است که تا حد ممکن با شواهد همبستگی داشته باشد. کلسترسازی همبستگی توسط بنسال و همکاران (Bansal et al., 2014) به جامعه نظری علوم کامپیوتر معرفی شده است. از آن زمان، این مسأله به‌طور وسیعی از منظر نظری مورد مطالعه قرار گرفته و همچنین در تعداد زیادی از برنامه‌های علم دیتا استفاده شده است.

عمومی‌ترین حالت کلسترسازی همبستگی توسط یک گراف $G = (V, W^+, W^-)$ ارائه می‌شود که در آن هر جوره رأس $(i,j) \in V \times V$ دارای دو وزن غیر منفی $w_{ij}^+ \in W^+ w_{ij}^- \in W^-$ است. این وزن‌ها نشان‌دهنده‌ی میزان شباهت و عدم شباهت i و j هستند. معمولاً، فقط یکی از این وزن‌ها برای هر جوره (i,j) صفر نیست تا نشان دهد که هر جوره رأس یا به طور کامل مشابه یا به طور کامل نامشابه هستند. توافق زمانی اتفاق می‌افتد که دو رأس مشابه در کنار هم کلستر شوند یا دو رأس نامشابه از هم جدا شوند. عدم توافق به وسیله‌ی دو رأس مشابه که از هم جدا شده اند یا دو رأس نامشابه که در کنار هم کلستر شده اند، تعریف می‌شود.

دو تابع هدف معمولی برای کلسترسازی همبستگی وجود دارد: ماکزیمم کردن وزن توافقات و مینیمم کردن وزن عدم توافقات. هنگام حداقل کردن عدم توافقات، قرار دادن رأس‌های i و j در یک کلستر با جریمه‌ی به مقدار w_{ij}^- همراه است، در حالی که جدا کردن آن‌ها جریمه‌ی به مقدار w_{ij}^+ دارد. بنابراین، هدف می‌تواند به صورت یک برنامه خطی صحیح (تام) (ILP) نوشته شود:

$$\begin{aligned} \text{minimize} \quad & \sum_{i<j} w_{ij}^+ x_{ij} + w_{ij}^- (1 - x_{ij}) \\ \text{subject to} \quad & x_{ij} \le x_{ik} + x_{jk} \text{ for all } i,j,k \\ & x_{ij} \in \{0,1\} \text{ for all } i,j \end{aligned} \tag{14}$$

در این فرمول‌بندی، x_{ij} نشان‌دهنده "فاصله" است: $x_{ij} = 0$ بیان می‌کند که رأس‌های i و j در کنار هم کلستر شده اند، در حالی که $x_{ij} = 1$ نشان‌دهنده این است که آن‌ها از هم جدا شده اند. شامل کردن محدودیت‌های نامساوی مثلثی $\left(x_{ij} \le x_{ik} + x_{jk}\right)$ اطمینان می‌دهد که خروجی یک کلستر معتبر از رأس‌ها را تعریف می‌کند. اولین مجموعه در هدف تمام جریمه‌های مربوط به قرار دادن رأس‌ها در کنار هم را محاسبه می‌کند و مجموعه دوم جریمه‌های ناشی از قرار دادن رأس‌ها در کلسترهای جداگانه را محاسبه می‌کند. یکی از ویژه گی‌های کلیدی کلسترسازی همبستگی این است که تعداد

کلسترهایی که باید تشکیل شود از قبل مشخص نشده است. بلکه، تعداد مناسب کلسترهایی که باید تشکیل شود به طور طبیعی با بهینه‌سازی (۱۲) به وجود می‌آید.

ماکزیمم کردن توافقات می‌تواند به صورت یک ILP نوشته شود که با یک ثابت مثبت از (۱۲) متفاوت است. به همین دلیل، هر دو مسأله زمانی که به طور دقیق بهینه‌سازی شوند معادل هستند. با این حال، از منظر تقریب‌ها، مینیمم کردن عدم توافقات به طور قابل توجهی چالش برانگیزتر است.

نسخه کامل و بدون وزن از یک کلسترسازی همبستگی یک گراف علامت دار را در نظر می‌گیرد که در آن هر دو رأس دارای یک ضلع مثبت یا یک ضلع منفی هستند. به طور معادل $(w_{ij}^{+}, w_{ij}^{-}) \in \{(0,1),(1,0)\}$، برای تمام $(i,j) \in V \times V$ است. بنسال و همکاران (Bansal et al., 2014) یک $PTAS$ برای ماکزیمم کردن توافقات و یک تقریب $O(1)$ برای مینیمم کردن توافقات ارائه کردند. چاریکار و همکاران (Charikar et al., 2017) اثبات کردند که حداقل کردن عدم توافقات APX-hard است؛ اما نسبت تقریب را به ۴ به‌طور قابل توجهی با گرد کردن یک تسهیل برنامه‌ریزی خطی از (۱۲) بهبود بخشیدند. نسبت تقریب برای حداقل کردن عدم توافقات بعداً به ۲,۵ توسط آیلون و همکاران (Hou et al., 2016) کاهش یافت، که همچنین یک تقریب بسیار سریع و زیبا با ضریب ۳ مبتنی بر تخنیکی به نام چرخش ارائه کردند. بهترین نسبت تقریب شناخته شده برای حداقل کردن عدم توافقات کمی‌کمتر از ۲,۰۶ است، که دلیل آن را می توان در مقاله (Chawla et al., 2015) دریافت. در گراف‌های با وزن‌های دلخواه، چندین گروه به‌طور همزمان اثبات کردند که مینیمم کردن عدم توافقات می‌تواند با گرد کردن یک تسهیل LP تا $O(\log n)$ تقریب زده شود (Charikar et al., 2017). برای ماکزیمم کردن توافقات، یک تقریب ساده نصفی با قرار دادن تمام رأس‌ها در یک کلستر یا جدا کردن هر رأس به کلستر خود حاصل می‌شود. بهترین تقریب شناخته شده برای ماکزیمم کردن توافقات ۰,۷۶۶۶ است که دلیل آن را می‌توان در (Christian, 2013) دریافت و تنها کمی بهتر از تقریب ۰,۷۶۶۴ است که چاریکار و همکاران در (Charikar et al., 2017) دریافتند. هر دو تقریب مبتنی بر گرد کردن یک تسهیل برنامه‌ریزی نیمه‌تعریف‌شده هستند.

مینیمم کردن عدم توافقات در گراف‌های با وزن دل‌خواه به‌عنوان معادل مینیمم چند قطع‌کننده شناخته می‌شود (Bonchi et al., 2015). این امر همچنین در موردی که همه رأس‌های که دارای یک ضلع مثبت، یا یک ضلع منفی، یا هیچ ضلع ای هستند، صدق می‌کند. این معادله بلافاصله نشان می‌دهد که به‌طور کلی، مینیمم کردن عدم توافقات UG-hard است تا در هر ضریب ثابتی قابل تقریب باشد (Chawla et al., 2015). علاوه بر این، امانوئل و فیات در (Damaschke, 2009) نشان دادند که

تسهیل برنامه‌ریزی خطی مسأله دارای یک شکاف یکپارچگی O(log n) است و اثبات کردند که تقریب O(log n) از طریق گرد کردن LP دقیق است.

نتایج نظری متعددی نیز برای انواع خاص کلسترسازی همبستگی اثبات شده است. بدون تلاش برای ارائه یک خلاصه جامع، اشاره می‌کنیم که این شامل نتایج برای گراف‌های دو قسمتی (Asteris et al., 2016)، گراف‌های رنگ‌آمیزی لبه (Bonchi et al., 2015) و هایپرگراف‌ها (Gleich et al., 2018) است. تحقیقات قبلی همچنین به محدودیت‌های مربوط به تعداد کلسترها (Anava et al., 2015)، توابع هدف جایگزین (Puleo & Milenkovic, 2018)، انواع وزنی خاص (Veldt et al., 2017) و الگوریتم‌های جریانی (Ahn et al., 2015) پرداخته‌اند.

کاربردها و الگوریتم‌های مقیاس‌پذیر

کلسترسازی همبستگی در دامنه‌های کاربردی متعددی از جمله تقسیم‌بندی تصویر (Kim et al., 2011 و Beier et al., 2014)، بیوانفورماتیک (Hou et al., 2016)، تشخیص لینک‌های چندزبانه (Anava et al., 2015) و تشخیص جامعه (Sharma & Singh, 2016) و (Wang et al., 2015) مورد استفاده قرار گرفته است. با این حال، بسیاری از الگوریتم‌های تقریبی نظری که توسعه یافته‌اند، در این موارد به‌طور خاص خوب عمل نمی‌کنند. برای شروع، بسیاری از بهترین نتایج نظری تنها به گونه‌هایی از مسأله مربوط می‌شوند که در عمل به ندرت پیش می‌آیند (مانند گراف‌های کامل و بدون وزن). چالش دیگری که وجود دارد این است که تسهیل‌های برنامه‌ریزی خطی و نیمه‌تعریف‌شده کلسترسازی همبستگی نیاز به حافظه بالایی دارند و حل آن‌ها در عمل دشوار است، حتی اگر از نظر نظری قابل حل در زمان چندجمله‌ای باشند. بنابراین، تمرکز مشترک در ادبیات بر ارائه تخنیک‌های مقیاس‌پذیر برای کلسترسازی همبستگی بوده است.

الگوریتم‌های مقیاس‌پذیر برای کلسترسازی همبستگی اشکال مختلفی دارند. چیریچتی و همکاران در (Chierichetti et al., 2014) و بعداً پان و همکاران در (Yang & Leskovec, 2015) نسخه‌های موازی الگوریتم محوری آیلون (Chung, 1997) را توسعه دادند. این الگوریتم‌ها با تضمین‌های تقریبی پیشینی همراه هستند؛ اما تنها به حالت کامل و بدون وزن اعمال می‌شوند. در طرف دیگر طیف، تخنیک‌های سریع ابتکاری که برای حالت عمومی وزنی معرفی شده‌اند (Beier et al., 2015) نشان می‌دهد که این تخنیک‌ها در برخی کاربردها عملکرد خوبی دارند؛ اما هیچ نوع تضمین تقریبی ارائه نمی‌دهند. بخش عمده‌یی از ادبیات مربوط به الگوریتم‌های مقیاس‌پذیر برای کلسترسازی همبستگی به‌طور خاص بر روی مسائل پراکنده متمرکز است که در آن‌ها اکثریت جوره‌های رأس ضلعی مشترک

ندارند. در این زمینه، کلسترسازی همبستگی وزنی اغلب به‌عنوان مشکل تقسیم‌بندی چندقطع‌کننده شناخته می‌شود. تعدادی تخنیک‌ها برای محاسبه حدود پایین برای هدف چندقطع‌کننده در عمل طراحی شده‌اند که راهی برای به‌دست آوردن تضمین‌های تقریبی پسینی فراهم می‌کنند (Lange et al., 2018) و (Swoboda & Andres, 2017). با این‌حال، این‌ها تسهیل LP مرسوم کلسترسازی همبستگی را حل نمی‌کنند و بنابراین، امکان پیاده‌سازی بهترین الگوریتم‌های تقریبی پیشینی را فراهم نمی‌آورند.

نتایج معادلت کلسترسازی گراف‌ها

ما دو تابع هدف را معادل می‌نامیم اگر زمانی که به‌طور بهینه حل شوند، خروجی یکسانی تولید کنند، حتی اگر از منظر تقریب‌ها یکسان نباشند. چندین هدف کلسترسازی گراف نشان داده شده‌اند که معادل یا حداقل به‌طور نزدیک مرتبط هستند. نیومن در نشان داد که بیشینه‌سازی مودولاریتی با یک پارامتر تفکیک معادل با ماکزیمم کردن یک تابع احتمال لوگاریتمی برای مودل بلوک تصادفی اصلاح‌شده بر اساس درجه است. چندین نویسنده به‌طور مستقل نشان دادند که نسخه نرمال‌شده‌ی از مودولاریتی با تعداد محدودی از کلسترها معادل با تعمیم k-way قطع‌کننده‌ی نرمال‌شده (۲,۷) است. دلوین و همکاران در همچنین رابطه‌ای بین مودولاریتی و قطع‌کننده نرمال‌شده مشاهده کردند که می‌توانند هر دو به‌عنوان موارد خاصی از هدف کلسترسازی پایدار در نظر گرفته شوند.

همان‌طور که در بخش اهداف تغییر یافته گراف کلستر شده اشاره شد، هدف مینیمم کردن اختلافات برای کلسترسازی همبستگی در گراف‌های بدون وزن معادل با ویرایش کلستر است. توابع هدف در واقع یکسان هستند: شمارش تعداد اضلاع‌هایی که باید به گراف اضافه یا حذف شوند تا آن را به یک گراف کلسترای تبدیل کنند، دقیقاً همانند شمارش تعداد اشتباهات ضلع منفی و اشتباهات ضلع مثبت در یک کلسترسازی از یک گراف امضا شده است. به‌طور مشابه، حذف کلستر می‌تواند به‌عنوان نسخه‌ی محدودشده از کلسترسازی همبستگی در نظر گرفته شود، که در آن فرد از ایجاد اشتباه در اضلاع‌های منفی منع شده است. سرانجام، آگاروال و کمپه در مشاهده کردند که تکنیک گرد کردن LP چریکار و همکاران در برای کلسترسازی همبستگی می‌تواند تطبیق داده شده و برای به‌دست آوردن حدود و راه‌حل‌های تقریبی برای مودولاریتی اعمال شود. اگرچه آگاروال و کمپه بر اثبات نتایج معادل تمرکز نکردند، این نتیجه نگاهی اولیه به رابطه بین دو هدف ارائه می‌دهد.

بحث و مناقشه

نتایج به‌دست‌آمده در این تحقیق به طور مستقیم با اهداف اصلی و سؤالات تحقیق ارتباط دارند. بررسی کلسترسازی گراف به‌عنوان یک مسأله بهینه‌سازی ترکیبی نشان داد که بسیاری از روش‌های موجود به دنبال به حداکثر رساندن تابع هدف نظیر مودولاریتی یا حداقل کردن هزینه‌های جداسازی کلسترها هستند. همچنین، نتایج به صورت واضح پیچیدگی محاسباتی الگوریتم‌ها را آشکار ساخت و ارتباط میان اهداف چندکلستری و اهداف مبتنی بر مودولاریتی را به تفصیل نشان داد. محدودیت‌های مودولاریتی از طریق مثال‌های عملی و نظری مشخص شد و در نهایت نتایج معادل در کلسترسازی گراف‌ها با استفاده از آنالیز ریاضی و تجربی، تفسیر گردید که در زیر به چند نکته آن اشاره شده است:

- یافته‌ها نشان می‌دهد که گراف‌ها با استفاده از رویکردهای بهینه‌سازی، مانند برنامه‌ریزی عدد تام و الگوریتم‌های فراابتکاری قابل کلسترسازی هستند. این مسأله به دلیل فضای جستجوی نمایی، بهینه‌سازی ترکیبی محسوب می‌شود.
- تجزیه‌وتحلیل‌ها نشان می‌دهد که بیشتر الگوریتم‌های کلسترسازی گراف در رده NP - سخت قرار می‌گیرند. این موضوع اهمیت توسعه الگوریتم‌های تقریبی و فراابتکاری را برجسته می‌کند.
- نتایج نشان داد که مودولاریتی، به‌رغم محبوبیت خود، گاهی اوقات نمی‌تواند الگوهای مقیاس کوچک را تشخیص دهد و این ضعف با اهداف چندکلستری تکمیل می‌شود. به‌طور خاص، مشخص شد که این معیار به خطای تفکیک مبتلا است، به این معنا که کلسترهای کوچک یا بسیار بزرگ را به‌درستی شناسایی نمی‌کند.
- این تحقیق تأیید کرد که برخی الگوریتم‌ها به نتایج معادلی می‌رسند که این امر نشان‌دهنده‌ی ساختارهای مشترک در گراف‌های تحلیل‌شده است.

یافته‌های این تحقیق همسو با مطالعاتی است که پیچیدگی محاسباتی کلسترسازی گراف را تأیید می‌کنند. به خصوص مانند کارهای "نیومن "و" فوتونیتو". مگر، در رابطه با محدودیت‌های مودولاریتی، نتایج این تحقیق تأکید بیشتری بر خطای تفکیک دارد که در برخی مطالعات گذشته کمتر به آن پرداخته شده بود. علاوه براین، تحلیل اهداف چندکلستری، نشان می‌دهد که این روش‌ها قابلیت کشف ساختارهای مقیاس کوچک‌تر را دارند که در بسیاری از مطالعات قبلی نادیده گرفته شده است.

نتیجه‌گیری

در این مقاله، مفاهیم بنیادی و استراتژی‌های اصلی در زمینه‌ی کلسترسازی گراف‌ها مورد بررسی قرار گرفتند. با پرداختن به اصول کلسترسازی، تعریف جامعه و نقش کلسترسازی به عنوان یک مسأله

بهینه‌سازی ترکیبی، چارچوبی جامع برای درک بهتر ساختارهای گراف و تشخیص جوامع ارائه شد. بررسی توابع هدف مبتنی بر مودولاریتی و اهداف تغییریافته‌ی کلسترسازی، همراه با تحلیل پیچیدگی محاسباتی این رویکردها، نه تنها درک عمیق‌تری از جنبه‌های نظری کلسترسازی گراف‌ها را فراهم کرد، بلکه به شفاف‌سازی روش‌های کاربردی برای حل این مسائل پیچیده نیز کمک کرد. همچنین، بررسی کاربردها و ارائه‌ی الگوریتم‌های مقیاس‌پذیر نشان می‌دهد که چگونه می‌توان از کلسترسازی گراف‌ها برای پروسس دیتاهای بزرگ و تحلیل شبکه‌های اجتماعی و بیولوژیکی استفاده کرد. نتایج معادل کلسترسازی گراف‌ها و پیش‌زمینه‌ی کلسترسازی همبستگی، ابزارهای مفیدی را برای بهبود دقت و کارایی در تشخیص ساختارهای جامعه در گراف‌های پیچیده ارائه می‌دهند.

به‌طور کلی، نتایج به‌دست‌آمده در این مقاله می‌توانند به‌عنوان یک مرجع کلیدی برای محققان و متخصصین این حوزه مفید واقع شوند و چارچوبی جامع برای تحقیق‌های آینده در کلسترسازی گراف‌ها و تحلیل ساختارهای شبکه‌ای فراهم کنند. این یافته‌ها همچنین نقش مهمی‌در توسعه‌ی الگوریتم‌های بهینه‌سازی ترکیبی برای تحلیل ساختارهای پیچیده و بهره‌برداری از آن‌ها در کاربردهای عملی ایفا خواهند کرد.

منابع

Abbe, E. (2018). Community detection and stochastic block models: Recent developments. *Journal of Machine Learning Research, 18*(177), 1-86. Retrieved from http://jmlr.org/papers/v18/16-480.html

Ahn, K. J., Cormode, G., Guha, S., McGregor, A., & Wirth, A. (2015). Correlation clustering in data streams. *In Proceedings of the 32nd International Conference on International Conference on Machine Learning, 37*, 2237-2246. Retrieved from http://dl.acm.org/citation.cfm?id=3045118.3045356.

Anava, Y., Avigdor-Elgrabli, N., & Gamzu, I. (2015). Improved theoretical and practical guarantees for chromatic correlation clustering. *In Proceedings of the 24th International Conference on World Wide Web, WWW, 15*, 55-65. https://doi.org/10.1145/2736277.2741629

Andersen, R., Chung, F., & Lang, K. (2006). Local graph partitioning using PageRank Vectors. *InProcedddings of the 47th Annual IEEE Symposium on Foundations of Computer Science*. Retrieved from http://www.math.ucsd.edu/~fan/wp/localpartition.pdf.

Arora, S., Rao, S., & Vaziani, U. (2009). Expander flows, geometric embeddings and graph partitioning. *Journal of the ACM (JACM), 56*(2). Retrieved from https://dl.acm.org/doi/abs/10.1145/1502793.1502794

Asteris, M., Kryillidis, A., Papailiopoulos, D., & Dimakis, A. G. (2016). Bipartite correlation clustering: Maximizing agreements. *In Proceeding of the 19th International Conference on Artificial Intelligence and Statistics (AISTATS), 51*. Retrieved from http://proceedings.mlr.press/v51/asteris16.html

Bader, D. A., Meyerhenke, H., Sanders, P., & Wagner, D. (2013). *Graph partitioning and graph clustering* (Vol. 588). American Mathematical Socity.

Bagon, S., & Galun, M. (2011). Large scale correlation clustering optimization. *arXiv preprint*. Retrieved from https://arxiv.org/abs/1112.2903

Bansal, N., Blum, A., & Chawla, S. (2014). Correlation clustering. *Machine Learining, 56*, 89-113. Retrieved from https://link.springer.com/article/10.1023/b:mach.0000033116.57574.95

Beier, T., Hamprecht, F. A., & Kappes, J. H. (2015). Fusion moves for correlation clustering. *In Proceeding of the IEEE Conference on Computer Vision and Pattern Recognition*, 3507-3516. Retrieved from http://openaccess.thecvf.com/content_cvpr_2015/html/Beier_Fusion_Moves_for_2015_CVPR_paper.html

Beier, T., Kroeger, T., Kappes, J. H., Kthe, U., & Hamprecht, F. A. (2014). Cut, glue, amp; cut: A fast, approximate solver for multicut partitioning. *In 2014 IEEE Conference on Computer Vision and Pattern Recongnition*, 73-80. https://doi.org/10.1109/CVPR.2014.17

Ben-Dor, A., Shamir, R., & Yakhini, Z. (1999). Clustering gene expression patterns. *Journal of computational biology*, 281-297. https://doi.org/https://doi.org/10.1089/106652799318274

Bhaskara, A., Daruki, S., & Venkatasubramanian, S. (2018). Sublinear algorithms for MAXCUT and correlation clustering. *In Proceeding International Conference on Automata, Logic and Programming*. Retrieved from https://arxiv.org/abs/1802.06992

Bhattacharya, A., & De, R. K. (2008). Divisive correlation clustering algorithm (DCCA) for grouping of genes: detecting varying patterns in expression profiles. *Bioinformatics*, 1359-1366. https://doi.org/10.1093/bioinformatics/btn133.

Bocker, S., & Baumbach, J. (2013). Cluster editing. *Springer Berlin Heidelberg*, 33-44. Retrieved from https://link.springer.com/chapter/10.1007/978-3-642-39053-1_5

Bolla, M. (2011). Penalized versions of the newman-girvan modularity and their relation to normalized cuts and k-means clustering. *phys*. https://doi.org/10.1103/PhysRevE.84.

Bonchi, F., Gionis, A., Gullo, F., Tsourakakis, C. E., & Ukkonen, A. (2015). Chromatic correlation clustering. *ACM Trans*, 1-34. https://doi.org/10.1145/2728170.

Bonomo, F., Duran, G., Napoli, A., & Valencia-Pabon, M. (2015b). Complexity of the cluster deletion problem on subclasses of chordal graphs. *Theoretical Computer Science*, 59-69. Retrieved from https://www.sciencedirect.com/science/article/pii/S0304397515005800

Boonomo, F., Duran, G., Napoli, A., & Valencia-Pabon, M. (2015a). A one-to-one correspondence between potential solution of the cluster deletion problem and the minimum sum coloring problem, and its application to p4-sparse graphs. *Information Processing Letters*, 600-603. Retrieved from https://lipn.fr/~valenciapabon/papers/cluster-del-chordal-v7.pdf

Buskirk, G., Fan, C., & Raichel, B. (2018). Metric violation distance: Hardness and approximation. *In Proceeding of the 29th Annual ACM-SIAm Symposium on Discrete Algorithms*, 196-209. Retrieved from https://epubs.siam.org/doi/abs/10.1137/1.9781611975031.14

Charikar, M., Gupta, N., & Schwartz, R. (2017). Local guarantees in graph cuts and clustering. *In International Conference on Integer Programming and Combinatorial Optimization*, 136-147. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-59250-3_12

Chawla, S., Makarychev, K., Schramm, T., & Yaroslavtsev, G. (2015). Near optimal lp rounding algorithm for correlationclustering on complete and complete k-partite graphs. *In Proceedings of the 47th Annual ACM on symposium on Theory of COmputing*, 219-228. Retrieved from https://dl.acm.org/doi/abs/10.1145/2746539.2746604

Chierichetti, F., Dalvi, N., & Kumar, R. (2014). Correlation clustering in mapreduce. *In Proceeding of the 20th ACM SIGKDD international conference on knowledge discovery and data mining*, 641-650. Retrieved from https://dl.acm.org/doi/abs/10.1145/2623330.2623743

Christian, S. (2013). High Quality Graph Partitioning. *PhD thesis, Karlsruhe Institute of Technology*. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=bb7389d0d8d4f8935e23b763c032eab20316b026

Chung, F. (1997). *Spectral Graph Theory*. American Mathematical Society.

Coleman, T., Saunderson, J., & Wirth, A. (2008). A local-seach 2-approximation for 2 correlation-clustering. *Springer Berlin Heidelberg*, 308-319. Retrieved from https://link.springer.com/chapter/10.1007/978-3-540-87744-8_26

Damaschke, P. (2009). Bounded-Degree Techniques Accelerate Some Parameterized Graph Algorithms. *Springer Berlin Heidelberg*, 98-109. https://doi.org/10.1007/978-3-642-11269-0 8.

Dau, P. L., Puleo, G., & Milenkovic, O. (2017). Motif clustering and overlapping clustering for social network analysis. *In IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, 1-9. https://doi.org/10.1109/INFOCOM.2017.8056956.

Delvenne, J., Yaliraki, S., & Barahona, M. (2010). Stability of graph communities across thime scales. *Proceedings of the National Academy of Sciences*, 12755-12760. Retrieved from https://www.pnas.org/doi/abs/10.1073/pnas.0903215107

Demain, E., Dinh, T., LI, X., & Thai, M. (2015). Network clustering via maximizing modularity: Approximation algorithms and theoretical limits. *In Proceeding of the 2015 IEEE International Xonference on Data Mining*, 101-110. https://doi.org/10.1016/j.tcs.2006.05.008.

Fan, C. (2009). A local graph partitioning algorithm using heat kernel pagerank. *Algorithms and Models for the Web-Graph*, 62-75. https://doi.org/10.1007/PL00012580.

Fortuato, S., & Hric, D. (2016). Community dtection in networks: A user guide. *Physics Reports*. https://doi.org/10.1016/j.physrep.2009.11.002.

Fukunaga, T. (2018). Lp-based pivoting algorithm for higher-order correlation clustering. *Springer Ingernational Publishing*. Retrieved from https://link.springer.com/article/10.1007/s10878-018-0354-y

Gao, Y., Hare, D., & Nastos, J. (2013). The cluster deletion problem for cographs. *Discrete Mathematics*, 2763-2771. Retrieved from https://www.mdpi.com/2504-2289/7/2/70

Gleich, D. F., Veldt, N., & Wirth, A. (2018). Correlation Clustering Generalized. *In 29th International Symposium on Algorithms and Computatiion*, 44. https://doi.org/10.4230/LIPIcs.ISAAC.2018.44.

Hou, J. P., Emad, A., Puleo, G. J., Ma, J., & Milenkovic, O. (2016). A new correlation clustering method for cancer mutation analysis. *Bioinformatics*, 3717-3728. https://doi.org/10.1093/bioinformatics/btw546.

Kim, S., Nowozin, S., Kohli, P., & Yoo, C. (2011). Higher-order correlation clustering for image segmentation. *Advances in Neural Information Processing Systems, 24*, 1530-1538. Retrieved from http://papers.nips.cc/paper/4406-higher-order-correlation-clustering-for-image-segmentation.pdf.

Kloumann, I. M., Ugander, J., & Kleinberg, J. (2017). Block models and personalized pagerank. *Proceedings of the National Academy of Sciences*, 33-38. https://doi.org/10.1073/pnas.1611275114.

Lange, J.-H., Karrenbauer, A., & Andres, B. (2018). Partial optimality and fast lower bounds for weighted correlation clustering. *Proceedings of the 35th International Conference on Machine Learning*, 2898-2907. Retrieved from proceedings.mlr.press/v80/lange18a.html.

Lu, L. F. (2002). Connected components in random graphs with given expected degree sequences. *Annuals of Combinatorics*, 125-145. https://doi.org/10.1090/cbms/092.

Neewman, M. (2016). Equivalence between modularity optimization and maximum likelihood methods for community detection. *physics*, 94-100. https://doi.org/10.1103/PhysRevE.94.052315.

Newman, M. E. (2013). Community detection and graph partitioning. *Europhysics Letters*, 103-105. Retrieved from stacks.iop.org/0295-5075/103/i=2/a=28003.

Pan, X., Papiliopoulos, D., Oymak, S., Recht, B., Ramchandran, K., & Jordan, M. (2015). Parallel correlation clustering on big graphs. *Advances in Neural Information Processing systems*, 82-90. Retrieved from papers.nips.cc/paper/5814-parallel-correlation-clustering-on-big-graphs.pdf.

Puleo, G. J., & Milenkovic, O. (2018). Correlation clustering and biclustering with locally bounded errors. *IEEE Transactions on Information Theory*, 4105-4119. https://doi.org/10.1109/TIT.2018.2819696.

Puleo, G., & Milenkovic, O. (2015). Correlation clustering with constrained cluster sizes and extended Weights bounds. *SIAM Journal on Optimization*, 1857-1872. https://doi.org/10.1137/140994198.

Schulz, C., Bayer, S. K., Hess, C., Steiger, C., Teichmann, M., Jacob, J., . . . Hayrapetyan, S. (2016). Graph partitioning and graph clustering in theory and practice. *Lecture notes at Institute for Theoretical Informatics Karlsruhe Institute of Technology*. Retrieved from https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=98ec106b2757cfcaa27673765b58f9a14a580e6e

Sharma, P., & Singh, M. (2016). Multi chromatics balls with relaxed criterion to detect larger communities in social networks. *Smart Trends in Information Technology and Computer Communication*, 196-203. Retrieved from https://link.springer.com/chapter/10.1007/978-981-10-3433-6_24

Swoboda, P., & Andres, B. (2017). A message passing algorithm for the minimum cost multicut problem. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition*, 4990-4999. https://doi.org/10.1109/CVPR.2017.530.

Veldt, N., Gleich, D., & Wirth, A. (2018). A correlation clustering framework for community detection. *In Proceeding of the 2018 WWW Conference*, 439-448. https://doi.org/10.1145/3178876.3186110.

Veldt, N., Wirth, A. I., & Gleich, D. F. (2017). Correlation clustrering with low-rank matrices. *In Proceedings og the 26th International Conference on World Wide Web*, 1025-1034. https://doi.org/10.1145/3038912.3052586.

Wang, D., Fountoulakis, K., Henzinger, M., Mahoney, M. W., & Rao, S. (2017). Capacity releasing diffusion for speed and locality. *Proceedings of the 34th International Conference on Machine Learning*, 3598-3607. Retrieved from proceedings.mlr.press/v70/wang17b.html.

Wang, Y., Huang, H., Feng, C., & Liu, Z. (2015). Community detection based on minimum-cut partitioning. *Web-Age Information Management*, 57-69. Retrieved from https://link.springer.com/chapter/10.1007/978-3-319-21042-1_5

Wirth, A., Veldt, N., Gleich, D., & Saunderson, J. (2018a). A projection method for metric-constrained optimization. *arXiv preprint arXiv*. Retrieved from https://arxiv.org/abs/1806.01678

Yang, J., & Leskovec, J. (2015). Defining and evaluationg network communities based on ground-truth. *Knowledge and Information Systems*, 181-213. https://doi.org/10.1007/s10115-013-0693-z.

Yu, L., & Ding, C. (2010). Network community discovery: Solving modularity clustering via normalized cut. *In Proceeding of the Eighth Workshop on Mining and Learning with Graphs*, 34-36. https://doi.org/10.1145/1830252.1830257.

Zhou, H., Li, J., Zhang, F., & Cui, Y. (2017). A graph clustering method for commuinty detection in complex network. *Physica A: Statistical Mechanics and its Applications*, 551-562. https://doi.org/10.1016/j.physa.2016.11.015.