A diagram of a research process

Description automatically generated

چکیده— مردم از شبکه‌های اجتماعی آنلاین مختلف برای اهداف مختلف استفاده می‌کنند. اطلاعات کاربر در هر شبکه اجتماعی معمولاً ناقص است. بنابراین، تطبیق کاربران در این شبکه‌های اجتماعی آنلاین چندگانه برای ارائه خدمات جدید و نیز درک بهتری از رفتار کاربر از اهمیت ویژه برخوردار است. تحقیقات اخیر نشان می‌دهد که ساختار گروه (*group structure*) به طور گسترده‌ای در شبکه‌های اجتماعی وجود دارد که در آن اعضا با هدف مشخص از همکاری پیروی کرده و در شبکه‌های اجتماعی آنلاین قدرت تأثیرگذاری بیشتری نسبت به فرد‌ها دارند. کارهای قبلی راهکارهای برجسته‌ای برای نقشه‌برداری افراد ارائه می‌دهند، اما توجه کافی به مطالعه گروه‌ها در سراسر شبکه‌های اجتماعی چندگانه ندارند. برای پر کردن این خلأ تحقیقاتی، ابتدا هدف ما ارائه یک روش مؤثر برای شناسایی گروه‌های مشابه در سراسر شبکه‌های اجتماعی چندگانه است. این روش عمدتاً شامل سه مرحله می‌شود که شامل شناسایی ساختار گروه براساس گذرهای تصادفی، استخراج ویژگی‌های مشابه و استنباط شباهت گروه با استفاده از مدل گرافی احتمالی است. ما الگوریتم خود را بر روی پنج نوع مختلف از شبکه‌های اجتماعی آنلاین ارزیابی کرده‌ایم. نتایج آزمایشی نشان می‌دهد که این روش عملکرد برجسته‌ای نسبت به روش‌های پیشرفته قبلی داشته‌است. علاوه بر این، نتیجه این مقاله راه را برای سیستم پیشنهاد، پیش‌بینی پیوند (*link prediction*) و انتشار اطلاعات در سایت‌ها به صورت قابلیت انتقال گشوده است.

شبکه‌های اجتماعی آنلاین نقش مهمی در زندگی روزانه مردم با انواع مختلفی از فرم‌ها مانند توییت‌ها، کسب و کار، به اشتراک گذاشتن موسیقی، خدمات بازی و وبلاگ‌ها ایفا می‌کنند. مردم برای اهداف مختلف حساب‌های شبکه‌های اجتماعی متعددی را می‌سازند. به عنوان مثال، کاربران به طور عمده با دوستان خود در شبکه اجتماعی فیس‌بوک ارتباط برقرار می‌کنند، گروه‌های علاقه‌مندی را از طریق استفاده از گوگل پلاس به اشتراک می‌گذارند و در لینکدین، ارتباطات شغلی خود را گسترش می‌دهند و... . یک نظرسنجی ساده توسط پژوهش‌گران نشان می‌دهد که میانگین تعداد شبکه‌های اجتماعی آنلاین که یک کاربر استفاده می‌کند، هشت است.

تحقیقات اخیر نشان می‌دهد که شناخت ارتباط کاربر در شبکه‌های اجتماعی آنلاین مختلف می‌تواند در راستای جلوگیری از گسترش اطلاعات اسپم و تشخیص مشتری‌های تقلبی کمک کند. بنابراین، کشف گروه‌های مشابه در سراسر شبکه‌های اجتماعی آنلاین اهمیت ویژه‌ای برای مطالعه رفتار گروهی دارد.

به عنوان مثال، بسیاری از شرکت‌ها مانند PeekYou و Spokeo خدمات جستجوی افراد بر اساس اطلاعات کاربران از شبکه‌های اجتماعی آنلاین را ارائه می‌دهند، علاوه بر این، ارائه دهندگان خدمات، مانند Yelp و Amazon، می‌توانند اسپمرها را با اتصال اکانت‌ها و تجزیه و تحلیل رفتارهای آنها در شبکه‌های اجتماعی آنلاین دیگر فیلتر کنند.

در این مقاله، گروه‌های مقایسه شده، از شبکه‌های اجتماعی آنلاین مختلفی هستند و گروه‌های مشابه باید کاربران مشترک و ساختار گروه مشابه را داشته باشند. ما در شکل ۱ به طور خلاصه مشکل خود را در کشف گروه‌های مشابه در سراسر شبکه‌های اجتماعی آنلاین چندگانه معرفی می‌کنیم. همانطور که در شکل 1 نشان داده شده است، کاربران اهداف مختلفی را برای خدمات وب در زندگی روزمره خود دارند. به منظور تحقق این اهداف، آنها در شبکه‌های اجتماعی آنلاین مختلفی مانند فیس‌بوک، گوگل پلاس و لینکدین فعالیت می‌کنند. در هر شبکه، ساختار گروه وجود دارد و ما باید این گروه‌ها را شناسایی کنیم. با محاسبه شباهت رابطه کاربران و محتوای اجتماعی در این دو گروه از دو شبکه مختلف در دنیای مجازی، قصد داریم بررسی کنیم که آیا آنها در دنیای واقعی هم گروه‌های یکسانی هستند یا خیر.

اطلاعات اسپم و نقدهای کاذب معمولاً از طریق گروه‌ها در شبکه‌های اجتماعی آنلاین منتشر می‌شوند. این گروه‌ها تأثیر قابل ملاحظه‌ای بر روی انتشار اطلاعات نسبت به کاربران فردی دارند. شناسایی و تحلیل این گروه‌ها در بین شبکه‌های اجتماعی مختلف دارای پیامدهای ارزشمندی برای تحقیقات درباره شبکه‌های اجتماعی آنلاین است. با پیوند دادن گروه‌های مشابه در شبکه‌های مختلف، می‌توانیم چالش اطلاعات ناقص را برطرف کرده و به درک جامع‌تری دست یابیم. همچنین، ویژگی‌های مختلف داخل یک گروه می‌تواند به تصفیه اطلاعات فردی در شبکه‌های اجتماعی مختلف کمک کند.

اگرچه حل این مسئله برجسته، امکان پذیر به نظر می‌رسد ولی با چالش‌های زیادی روبه‌رو است. اولاً با گسترش روز افزون حجم داده در رسانای جمعی، امکان بررسی همه داده‌ها وجود ندارد. بنابراین مشکل کمبود اطلاعات برای برخی کاربران اجتناب ناپذیر است. این مانع به چند دلیل ایجاد می‌شود. برای مثال کاربر اجازه دسترسی به اطلاعاتش را نمی‌دهد یا به دلیل مجموعه‌ای از عوامل اطلاعاتش به درستی ارسال نمی‌شود (مثلا سرعت اینترنت ایران و فیلترینگ می‌تونه یکی از این عوامل باشه😊). ثانیاً پیچیدگی محاسباتی بالا همواره چالش اساسی در تحلیل کلان داده‌ها (big data) به خصوص برای گراف‌های بزرگ است که اجرای موازی آن‌ها می‌تواند دشوار باشد. کشف ساختار گروه‌ها یک مسئله خوشه بندی رئوس در گراف‌هاست که به عنوان یک مسئله NP-Complete اثبات شده‌است (این اصطلاح به یک سری مسئله تصمیم‌گیری اطلاق میشه که هیچ الگوریتم شناخته‌شده قابل اجرایی براشون وجود نداره. تا اونجایی که می‌دونیم زمان اجرای بسیار زیاد باعث حل نشدن این مسائل میشه). علاوه بر این برای محاسبه شباهت دو گروه در شبکه‌های اجتماعی آنلاین مختلف منابع محاسباتی زیادی مورد نیاز است. در اینجا تحلیلی ساده ارائه می‌دهیم: G1 و G2 دو شبکه را نشان می‌دهند و n1 و n2 تعداد کاربران در این شبکه‌ها هستند. روی کاغذ پیچیدگی زمانی محاسبه مقدار شباهت، برای همه جفت گروه‌ها بدون شک از O(n1n2) کوچکتر نیست! (برای اطلاعات بیشتر درباره big O notation و پیچیدگی زمانی می‌تونید به [این لینک](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) مراجعه کنید.) ثالثاً، پیدا کردن مبنای پیش فرض (ground truth) برای برچسب گذاری روی کاربران شبکه‌های اجتماعی یکی از چالش برانگیزترین موانع در مسیر بررسی کاربران است.

سخن مترجم - یه مثال برای درک مفهوم مبنای پیش فرض:‌

فرض کنید می‌خواهید الگوریتم کامپیوتری برای شناسایی گربه‌ها در تصاویر توسعه دهید. برای آموزش و آزمون الگوریتم، نیاز به یک مجموعه داده از تصاویر با برچسب حقیقت مبنا در مورد وجود یا عدم وجود گربه دارید.

برای تشکیل حقیقت مبنا، مجموعه‌ای از تصاویر را جمع‌آوری می‌کنید و به صورت دستی به تمام تصاویر مراجعه کرده و با دقت حضور یا عدم حضور گربه را شناسایی و مشخص می‌کنید. این تصاویر با برچسب‌های حقیقت مبنا، مجموعه داده حقیقت مبنا شما می‌شوند. سپس، هنگامی که الگوریتم را روی تصاویر جدید غیرمشهود اجرا می‌کنید، می‌توانید پیش‌بینی‌ها را با برچسب‌های حقیقت مبنا مقایسه کرده و صحت عملکرد آن را ارزیابی کنید.

به طور خلاصه، حقیقت مبنا پایه ایمن و قابل اعتمادی برای ارزیابی صحت و کیفیت داده‌ها، الگوریتم‌ها، مدل‌ها یا پیش‌بینی‌ها در انواع زمینه‌های تحقیقاتی، یادگیری ماشین و تحلیل داده‌ها فراهم می‌کند.

برای توصیف این مسئله، ابتدا مجموعه شبکه‌های اجتماعی آنلاین چندگانه را با نام Gs در نظر می‌گیریم، که توسط شبکه‌های اجتماعی مختلف G1، G2،...، Gk ساخته می‌شوند.همانطور که در شکل 2 نشان داده شده است, هر شبکه شامل محتوای کاربران و روابط آنها است. فرض کنید g1 و g2 به ترتیب دو گروه از دو شبکه اجتماعی آنلاین مختلف Gi و Gj باشند که در Gs قرار دارند. هدف ما در این حالت این است که با مقایسه کاربران در گروه و ساختار گروه، الگوریتمی بیابیم به طوری که بتواند تشخیص دهد آیا g1 و g2 از کاربران یکسان با روابط مشابهی تشکیل شده‌اند یا خیر. ما این مسئله را به دو دسته تقسیم می‌کنیم: در دسته اول ما داده‌ هایی که حاکی از روابط میان گروهی باشد، نداریم. در این حالت تنها داده‌های در دسترس،‌ اطلاعات پروفایل و چیدمان روابط کاربران (Topology) در هر گروه است. در دسته دوم ما اطلاعات میان گروهی و درون گروهی کافی را در اختیار داریم. طبیعتاً چالش اصلی درباره دسته اول است که در این مقاله به آن‌ها پرداخته می شود(شکل ۲). خوب است یادآوری کنیم که در اینجا واژه‌های گراف و شبکه کاربرد یکسانی دارند.

متدولوژی: پس از وارد کردن هر شبکه‌ اجتماعی آنلاین، ابتدا از گذرهای تصادفی (*Random Walks*) برای کشف گروه‌ها در هر شبکه استفاده می‌کنیم. سپس، شباهت هر زوج گروه از شبکه‌های مختلف را از دیدگاه‌های متفاوت، از جمله پروفایل گروه و چیدمان روابط، استخراج می‌کنیم. مدل گرافیِ احتمالی که بر اساس مجموعه حقیقت‌های ساخته شده است، برای استنباط اینکه آیا دو گروه مشابه هستند یا خیر استفاده می‌شود.

هدف ما شناسایی گروه‌ها در شبکه‌ها است. یک گروه می‌تواند از حداقل دو کاربر نزدیک که با یکدیگر همکاری دارند یا به نوعی با هم دوست هستند،‌ تشکیل شود. در این مطالعه، از روش [پیمایش‌های تصادفی](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) استفاده می‌شود تا ساختار گروه‌ها را در چندین شبکه پیدا کنیم. پیمایش‌های تصادفی باعث می‌شوند که به زیرگراف‌های مجاورِ متصل بیفتیم،که به عبارتی همان ساختار گروه است که ما می‌خواهیم کشف کنیم. با استفاده از این روش، می‌توانیم به طرز موثری گروه‌ها را شناسایی کنیم.

بهره‌گیری از [تئوری زنجیره مارکوف](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) و [طیف گراف](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) نیز تأییدی ریاضی را برای این روش ارائه می‌دهد.

در استفاده از روش پیمایش تصادفی، با در نظر گرفتن یال‌هایی که پیمایش‌ها طی می‌کنند و همچنین یال‌هایی که بین هر زوج رئوس در مسیر وجود دارد، ما شاخصی برای نشان دادن نزدیکی رئوس پیشنهاد می‌دهیم.

برای تشریح ساده‌تر مطلب، یک نمونه گرافی را در نظر می‌گیریم که شامل دو زیرگراف S1 و S2 است. مشاهده می‌کنیم که تعداد مسیرهای خاص در S1 بیشتر از S2 است. از این تفاوت الهام گرفته و یک شاخص را برای نزدیکی رئوس پیشنهاد می‌دهیم.

با پیمایش تصادفی، مسیرهایی که در زیرگراف کامل وجود دارند بیشتری احتمال دارند توسط پیمایش‌گر طی شوند نسبت به مسیرهای در زیرگراف دور. با استفاده از این احتمالات، می‌توانیم احتمال محاسبه یک یال در زیرگراف‌ها را محاسبه کنیم. از تفاوت احتمال محاسبه یال‌ها در زیرگراف‌های متصل به هم و زیرگراف‌های با اتصال پراکنده بهره می‌گیریم تا به شناسایی گروه‌ها بپردازیم.

با استفاده از این روش، می‌توانیم در هر شبکه گروه‌ها را شناسایی کنیم. ابتدا یک راس اولیه را به صورت تصادفی انتخاب کرده و سپس با استفاده از پیمایش تصادفی پیمودن مسیر، اعضای گروه آن را کشف می‌کنیم.

عملکرد یک گروه معمولا براساس کاربرانی که دوستان بیشتری دارند تعیین می‌شود. ضریب باشگاه غنی (rc) نسبت تعداد یال ها به رئوس با درجه بزرگتر از k است. در واقع این ضریب نشان دهنده رابطه بین کابران رهبری کننده در یک گروه است.

در این معادله Nk تعداد رئوس با درجه بیشتر از k است و Ek تعداد یال های بین این رئوس است. از آنجا که باشگاه غنی دو گروه ممکن است تفاوت های زیادی در تعداد درجه‌هایشان داشته باشند، مقدار k برای هر گروه 0.1 \* dmax در نظر گرفته می‌شود که dmax راسی با بیشترین درجه بین دو گروه است. اگر ۲ گروه، رهبرانی با روابط و ساختار یکسان داشتند، در این صورت باید ضریب باشگاه غنی آن دو گروه نیز یکسان باشد حتی با وجود تفاوت زیاد در تعداد درجه ها. اگر φ1 و φ2 به ترتیب ضریب دو گروه باشند، شباهت باشگاه غنی دو گروه را بر اساس این عبارت می‌سنجیم:

ضریب خوشه‌بندی که آن را با CC نمایش می‌دهیم، نشان دهنده نزدیکی کانکشن‌های هر کاربر است. همچنین این شاخص برای اندازه گیری شباهت میان گروه‌ها نیز استفاده می‌شود.

در این عبارت g نشان دهنده یک گروه، Tv نشان دهنده تعداد مثلث های موجود با راس v، dv نشان دهنده درجه v و ng تعداد رئوس در گروه g است. همچنین شباهت دو گروه می‌تواند با استفاده از فرمول بالا نیز محاسبه شود.

ج) استنباط بر اساس مدل گرافیکی احتمالاتی

روند کار مدل استنباطی ما در تصویر شماره هفت نشان داده شده است که می‌توان از آن برای فهمیدن اینکه آیا دو گروه مشابه هستند یا نه استفاده کرد. ما شباهت های گفته شده در بخش III-B را به عنوان ورودی دریافت می‌کنیم و با بیشینه کردن مقدار BDeu، رابطه بین گروه‌های مشابه که با گراف های احتمالاتی نشان داده شده‌اند را پیدا می‌کنیم. این گراف ها برای ساختن مدل شبکه ‌Bayesian برای پیش‌بینی احتمال شباهت دو گروه استفاده می‌شوند.

مدل گرافیکی احتمالاتی، از یک گراف استفاده می‌کند تا پیش نیاز های بین متغیر های تصادفی را بیان کند. در یک گراف احتمالاتی، راس ها نشان دهنده متغیرها و یال ها نشان دهنده رابطه بین رئوس است. تصویر شماره ۸ همین موضوع را نشان می‌دهد. یالی در این تصویر وجود دارد که راس Vp را به راس Vveo متصل می‌کند. این یال نشان دهنده این موضوع است که simp یک تاثیر علّی برای یال simveo است.

1. مدل شبکه بیزی (Bayesian): در مسئله ما اطلاعات مفید، خیلی پراکنده هستند و همچنین بسیاری از داده‌ها ناقص هستند. بنابراین داده‌های آموزشی محدود است و بسیاری از آن ها نامشخص هستند.  
     
   **مدل شبکه بیزی**: در مسئله ما، اطلاعات مفید بسیار پراکنده هستند و بسیاری از داده‌ها ناقص هستند. بنابراین، اندازه مجموعه داده‌های آموزشی محدود است و بسیاری از آن‌ها نامشخص هستند. مدل شبکه بیزی با استفاده از چندین گراف احتمالی می‌تواند به نتیجه‌گیری‌هایی برسد که با وجود اطلاعات ناقص خطا نداشته باشد. علاوه بر این، رویکرد شبکه بیزی می‌تواند تاثیرگذاری متغیرها را به خوبی کشف کند و استنباط‌های خوبی برای مدل‌سازی عدم قطعیت بر اساس چندین گراف احتمالی انجام دهد. بنابراین، رویکرد شبکه بیزی نیازهای ما را برآورده می‌کند و می‌تواند در مسئله استنباط ما استفاده شود.

بخش چهارم. ارزیابی

در این بخش ما جزئیات آزمایش خود را بیان می‌کنیم و به مقایسه بین روش های ما و جدیدترین روش ها می‌پردازیم. ما پیش‌بینی یک پیوند را به عنوان موضوع مطالعه انتخاب کردیم تا پیشرفت روش های ارائه شده را بررسی کنیم.

الف) تنظیمات آزمایش

این اطلاعات از COSNET به دست آمده‌است که یک مقاله است با محوریت کانکت شدن کاربران مشابه در شبکه های اجتماعی مختلف. این مجموعه اطلاعات روابط بین کاربران و همچنین پروفایل های کاربران را در پنج رسانه اجتماعی فراهم می‌کند. همانطور که در جدول شماره یک آمده است، تعداد کاربران زیاد است اما پروفایل‌هایی که اطلاعات مهم را فراهم می‌کنند بسیار کمتر و بسیار پراکنده هستند. فرضیات ما برای جفت کاربران از COSNET تامیین شده است. اما به جای جفت کردن کاربران، ما نیاز داریم گروه‌ها را جفت کنیم. در اینجا ما ایده اصلی را برای تبدیل جفت کاربران به جفت گروه‎‌ها بیان می‌کنیم. تحقیقات نشان می‌دهد که هر گروه معمولا دارای تعداد کمی کاربر است با درجات بالا و تعداد زیادی کاربر با درجات کم. در نتیجه ما کاربران هر گروه را به دو دسته تقسیم می‌کنیم. کاربرانی با تعداد درجات بالا که تعداد درجه آنها جزو 20 درصد اول کاربران است و بقیه 80 درصد کاربران که نسبتا درجه کمتری دارند. برای آموزش مدل، ما بخشی از داده‌ها را به صورت تصادفی انتخاب می‌کنیم. اندازه این بخش به تعداد کل نمونه ها در مجموعه داده و فاصله نمونه گیری بستگی دارد. بقیه داده ها برای آزمون مدل استفاده می‌شوند. ما با استفاده از داده‌های آموزشی، گراف‌های احتمالی را به دست می‌آوریم و از این گراف‌ها برای محاسبه احتمال مشابه بودن دو گروه استفاده می‌کنیم.

برای بررسی اثر این دو پارامتر یعنی اندازه نمونه معادل و فاصله نمونه گیری، آنها را در دو مجموعه مختلف تنظیم می‌کنیم. از هر مجموعه پنج گراف احتمالی برتر را انتخاب می‌کنیم تا مدل بیزی را بسازیم. همه این آزمایش های گفته شده روی یک سرور با پردازنده چهار هسته ای و 16 گیگابایت رم انجام شده است.

برای بررسی و ارزیابی نتایج به دست آمده، آنها را با سه معیار دیگر مقایسه میکنیم:

ب) نتایج آزمایش

ما جزئیات رقبای خود و همچنین نتیجه مقایسه را بیان می‌کنیم. ما همچینن از یک مطالعه موردی برای نشان دادن اهمیت نتیجه خود برای بهبود عملکرد برنامه های دیگر استفاده می‌کنیم.

1. رقبا:

استخراج داده‌های شبکه‌های اجتماعی در سال‌های اخیر بسیار موضوع داغی بوده‌است. چندین مقاله مانند این مقاله وجود دارد که تلاش می‌کنند کاربران مشابه را در شبکه‌های مختلف پیدا کنند. ما روش پیاده سازی برخی از آنها را معرفی می‌کنیم:

COSNET :‌آنها از پروفایل کاربران و کانکشن هر کاربر برای محاسبه مقدار شباهت کاربران از یک مدل مبتنی بر انرژی استفاده می‌کنند.

PathSim : آنها یک روش مبتنی بر مسیرمتا پیشنهاد می‌کنند تا از یک مسیر متا برای پیدا کردن محتوای کاربران در شبکه‌های اجتماعی استفاده می‌کنند و درجه شباهت کاربران را با محاسبه مقدار شباهت مسیرهای متای آنها استنتاج می‌کنند.

HeteSim : این روش مشابه روش قبلی است. آنها نیز مسیرهای متا را در نظر می‌گیرند و شباهت کاربران را بر اساس این مسیرها به دست می‌آورند.

از آنجا که این کارهای قبلی عموما بر روی پیدا کردن کاربران مشابه تمرکز کرده‌اند، در نتیجه نمی‌توانند گروه‌ها را به طور مستقیم تشخیص دهند. حال گروه‌ها را به عنوان یک کاربر در نظر می‌گیریم به طوری که پروفایل آن را پروفایل تمامی کاربران داخل گروه در نظر می‌گیریم. و در نهایت این روش‌ها را رو شبکه ها اعمال می‌کنیم و گروه‌های مشابه را که با کاربران جایگزین شده اند پیدا می‌کنیم.

1. نتایج شناسایی گروه‌ها  
   ما روش خود را با رقبا در پنج شبکه اجتماعی آنلاین مختلف مقایسه کردیم. مقادیر به دست آمده که میانگین داده‌ها هستند در شکل 9 قابل مشاهده است. با توجه به این شکل، می‌توانیم ببینیم که عملکرد الگوریتم ما در تمامی سه معیار از رقبا بهتر است. بهبودهای دقت، بازخوانی و معیار F1 بین ۲۰.۸٪ تا ۴۹.۱٪، ۱۰.۶٪ تا ۴۴.۷٪ و ۱۷.۹٪ تا ۴۷.۶٪ نسبت به الگوریتم‌های مقایسه‌ای بوده است. بهبودهای کلی بین ۳۱.۴٪ تا ۴۴.۳٪ در دقت، ۱۷.۳٪ تا ۲۶.۳٪ در بازخوانی و ۲۵.۹٪ تا ۳۱.۶٪ در معیار F1 بوده است. نتایج بهتر، اعتبار مدل ما را تأیید می‌کنند.
2. تحلیل ویژگی‌ها:

برای نشان دادن مشارکت شش ویژگی گفته شده در بخش III-B، ما آن‌ها را با اندازه‌گیری افزایش اطلاعات می‌سنجیم که نتیجه آن در شکل 10 قابل مشاهده است. شکل 10 نشان می‌دهدد که شباهت پروفایل بیشترین اهمیت را در نتیجه پایانی دارد که تا 0.607 درصد به افزایش اطلاعات کمک می‌کند.بعد از آن شباهت راس و یال ها قرار دارد که مقدار آن 0.501 است. این به این معناست که گروه های مشابه، معمولا کاربران مشابهی با پروفایل و روابط مشابه در شبکه های اجتماعی مختلف دارند. شباهت تعداد درجه (کانکشن) هر کاربر کمترین تاثیر را دارد. زیرا مقیاس شبکه‌های اجتماعی، متفاوت است و این عدد معمولا در شبکه‌های اجتماعی، معنای متفاوتی دارد.

ج) بهبود سیستم

برای بررسی اینکه آیا یافته‌های ما می‌توانند در برنامه‌های دیگر نیز مفید باشد، مساله پیش‌بینی لینک را بررسی می‌کنیم. منظور از پیش‌بینی لینک این است که پیش‌بینی کنیم آیا در آینده دو کاربر به هم متصل می‌شوند یا نه. این یک مساله معروف در تحقیقات شبکه‌های اجتماعی است. ما از روش Adamic-Adar به عنوان الگوریتم پایه استفاده می‌کنیم. این روش به همسایگان مشترک بین دو نفر در شبکه اجتماعی نگاه می‌کند و یک شاخص برای پیش‌بینی ارتباط بین آنها می‌سازد.

فرض کنید ما چندین گروه مشابه داریم که دو نفر u و v در آن‌ها حضور دارند. ما از این اطلاعات برای بهبود پیش‌بینی لینک استفاده می‌کنیم. تابعی ساده برای این کار داریم:

* اگر u یا v در گروه نباشند: مقدار 0
* اگر u و v در گروه باشند ولی ارتباطی نداشته باشند: مقدار 0.5
* اگر u و v در گروه باشند و ارتباط داشته باشند: مقدار 1

با استفاده از این تابع، شاخص Adamic-Adar برای گروه‌های مشابه بهبود می‌یابد. اگر u و v در چندین گروه مشابه باشند، شاخص آن‌ها افزایش می‌یابد. نتایج نشان می‌دهد که وقتی الگوریتم جدید لینک‌ها را پیش‌بینی می‌کند، از اطلاعات گروه‌های مشابه برای بهبود عملکرد خود استفاده می‌کند. در شبکه‌های مختلف، بهبود الگوریتم حدود 0.25 است. همچنین، با افزایش تعداد شبکه‌هایی که از آن‌ها استفاده می‌کنیم، دقت آزمایش افزایش می‌یابد. وقتی تعداد شبکه‌ها به 5 می‌رسد، دقت آزمایش به حالت ثابتی می‌رسند.

مقیاس‌پذیری

در این مقاله، فقط پروفایل کاربران و روابط بین آنها را در نظر می‌گیریم. در صورت دسترسی، اطلاعات دیگر مانند زمان ورود به برنامه، اطلاعات سفر، محتوای عمومی و اشتراک گذاری لحظه‌ها نیز می‌توانند برای بهبود عملکرد روش ما مورد استفاده قرار گیرند. زمان اجرای محاسبات برای گراف‌های بزرگ همیشه یک چالش است. شناسایی گروه‌ها و محاسبه شباهت روابط بین کاربران بیشترین زمان اجرا را اشغال می‌کنند. یکی از محبوب‌ترین روش‌ها برای کاهش زمان اجرا، محاسبات موازی است که می‌تواند در روش ما نیز استفاده می‌شود. ما می‌توانیم از چندین حرکت تصادفی مستقل برای بررسی ساختار گراف استفاده کنیم. در این حالت، محاسبات این بخش می‌تواند به صورت موازی انجام شود، همانطور که در کار قبلی ما [14] بیان شده است. از آنجایی که محاسبه شباهت کاربران مستقل است، موازی‌سازی آن‌ها دشوار نیست. با تقسیم محاسبات به بخش‌های کوچکتر و ارسال آنها به تعدادی کامپیوتر، هزینه زمانی بسیار کمتر می‌شود که برای پردازش رسانه‌های اجتماعی بزرگ مناسب است. زمان اجرا بر روی تعداد مختلف کامپیوتر در شکل 12 نشان داده شده است همانطور که در نمودار مشخص است، زمان اجرا با افزایش منابع محاسباتی به طور قابل توجهی کاهش می‌یابد و همچنین این روش توانایی مقابله با بیگ دیتاها (big data) را نشان می‌دهد.

**نتیجه‌گیری**

ما یک مسئله مهم را مطرح می‌کنیم و فرمول‌بندی می‌کنیم؛ کشف گروه‌های مشابه در چندین شبکه اجتماعی، که می‌تواند عملکرد برنامه‌های دیگر را بهبود ببخشد. برای حل این مسئله، ما یک روش موثر پیشنهاد می‌دهیم: ابتدا، از طریق پیمایش‌های تصادفی، ساختار گروه در هر شبکه را به دست می‌آوریم، سپس ویژگی‌های شباهت گروه را از زوایای مختلف استخراج می‌کنیم و بر اساس مدل گرافی احتمالی، نهایتاً تشخیص می‌دهیم که آیا دو گروه مشابه هستند یا خیر.  
برای تحقیق و پژوهش بیشتر در زمینه شبکه‌های اجتماعی می‌توانید به دو عنوان تحقیقاتی*Social Media Mining* و *Community Detection* که مرتبط با موضوع همین مقاله هستند، مراجعه کنید.