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 نشان دهنده آن است که گروه‌ها نسبت به افراد قدرت تاثیر گذاری بیشتری در شبکه‌های اجتماعی دارند. 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 عبارتند از شناسایی ساختار گروه براساس گذرهای تصادفی، استخراج ویژگی‌های مشابه و استنباط شباهت گروه با استفاده از مدل گرافی احتمالی. 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.

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.

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. کاربران اهداف مختلفی را برای خدمات وب در زندگی روزمره خود دارند. 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. 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 رسانه‌های 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 می‌خواهیم الگوریتم کامپیوتری برای شناسایی گربه‌ها در تصاویر توسعه دهیم. برای آموزش و آزمون الگوریتم، نیاز به یک مجموعه داده از تصاویر با برچسب حقیقت مبنا در مورد وجود یا عدم وجود گربه داریم. برای تشکیل حقیقت مبنا، مجموعه‌ای از تصاویر را جمع‌آوری می‌کنیم و به صورت دستی به تمام تصاویر مراجعه کرده و با دقت حضور یا عدم حضور گربه را شناسایی و مشخص می‌کنیم. این تصاویر برچسب گذاری‌شده همان مجموعه حقیقت مبنای ما را تشکیل می‌دهند. سپس، هنگامی که الگوریتم را روی تصاویر جدید اجرا می‌کنیم، می‌توانیم با استفاده از حقایق مبنا تشخیص دهیم که گربه داریم یا نه. 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.

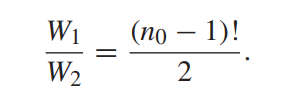
پس از وارد کردن هر شبکه‌ اجتماعی آنلاین، ابتدا از گذرهای تصادفی (*Random Walks*) برای کشف گروه‌ها در هر شبکه استفاده می‌کنیم. سپس، شباهت هر زوج گروه از شبکه‌های مختلف را از دیدگاه‌های متفاوت، از جمله پروفایل گروه و چیدمان روابط، استخراج می‌کنیم. هدف ما شناسایی گروه‌ها در شبکه‌ها است. یک گروه می‌تواند 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.

الف) شناسایی گروه‌ها در شبکه‌ها:  
برای کشف ساختار گراف در زمینه نمونه‌گیری و برآورد شبکه‌ها، از گذر تصادفی استفاده می‌شود. چون گذر تصادفی به‌راحتی در زیرگراف‌های متراکم‌ (*closely connected*) گرفتار می‌شود، که این همان ساختار گروهی است که ما می‌خواهیم کشف کنیم. با استفاده از این ویژگی گذر تصادفی، می‌توانیم گروه‌ها را به‌طور مؤثری شناسایی کنیم. علاوه بر این، بهره‌گیری از [تئوری زنجیره مارکوف](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) و [طیفی گراف](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) نیز تأییدی ریاضی را برای این روش ارائه می‌دهد.

شکل 4 یک مثال از گرافی با زیرگراف‌های S1 و S2 است که در آن S1 مسیرهای بیشتری نسبت به S2 دارد. برای ساده‌سازی تحلیل، فقط یک نوع مسیر خاص (سه یال و چهار راس متفاوت) را در نظر می‌گیریم. S1 دارای 24 مسیر متفاوت از این نوع خاص است، در حالی که S2 فقط 8 مسیر متفاوت از این نوع دارد. با الهام از این تفاوت، ما یک شاخص برای نشان دادن نزدیکی رئوس پیشنهاد می‌کنیم:

فرض می‌کنیم W1 تعداد این مسیرهای خاص در زیرگراف کاملی است که در آن هر گره به تمام گره‌های دیگر متصل است (*Complete Subgraph*). از سوی دیگر W2 نشان دهنده تعداد این مسیرهای خاص در زیرگراف

چرخه‌ای است که در آن هر گره به ۲ گره متصل است (*Cycle Subgraph*). همچنین n0 نشان دهنده تعداد راس‌ها در این دو زیرگراف است. نتیجه بررسی‌ها نشان می‌دهد که W1 = n0! و W2 = 2n0 است که در نهایت داریم:A black and white math equation

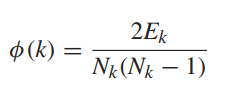
Description automatically generated with medium confidence

با استفاده از [مجموعه‌ای از محاسبات](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper)، از رابطه بالا استنتاج می‌شود که تعداد مسیرهای خاص در زیرگراف کامل بسیار بیشتر از زیرگراف چرخه‌ای است. در نتیجه احتمال طی کردن یک مسیر خاص توسط گذر تصادفی در زیرگراف کامل بیشتر از زیرگراف چرخه‌ای است. با این توصیفات، وزن یک یال به تعداد مسیرهای خاصی که از آن عبور می‌کنند و احتمال طی شدن آن مسیرها توسط گذر تصادفی بستگی دارد. در نتیجه، یال‌هایی که در زیرگراف‌های کامل‌تر (با ارتباطات نزدیک‌تر) هستند، وزن بیشتری نسبت به یال‌هایی که در زیرگراف‌های چرخه‌ای (با ارتباطات کمتر) هستند، خواهند داشت. این روش کمک می‌کند تا گروه‌های متراکم‌تر در شبکه‌ها شناسایی شوند.

ب) شباهت گروه‌ها در چندین شبکه اجتماعی آنلاین  
۱- شباهت پروفایل: اطلاعات پروفایل کاربر توصیف پایه‌ای راجع به یک کاربر ارائه می‌دهد. برای مثال، نام کاربری، اطلاعات مکان، صفحه اصلی پروفایل کاربر، محل تولد، سن کاربر، وضعیت تأهل، وضعیت شغلی و... . تحقیقات قبلی نشان داده است که کاربران مشابه در چندین شبکه اجتماعی آنلاین باید پروفایل مشابهی را به اشتراک بگذارند. به جای بهره‌برداری از همه اطلاعات پروفایل، ما از سه عامل نام کاربری، مکان و صفحه اصلی پروفایل استفاده می‌کنیم. زیرا این توضیحات معمولاً شخصیت‌های آشکار شخص را دارند که او را از سایر کاربران متمایز می‌کنند. علاوه بر این، این موارد معمولاً در چندین شبکه اجتماعی شبیه به هم هستند که به ما در کشف آن‌ها در شبکه‌های اجتماعی آنلاین مختلف کمک می‌کند. ما این رشته‌ها را برای یک کاربر به صورت یک کیسه کلمات تقسیم می‌کنیم و از الگوریتم [TF-IDF](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) برای پردازش کیسه کلمات و بدست آوردن بردار وزنی آن استفاده می‌کنیم. ما می‌خواهیم از بردارهای وزنی Wg1 و Wg2 استفاده کنیم تا نتایج TF-IDF را برای g1 و g2 نشان دهیم. در نهایت، عبارت زیر نشان دهنده کسینوس زاویه بین دو بردار است و هر چه این مقدار به یک نزدیکتر باشد میزان شباهت بین دو پروفایل بیشتر است.

A graph of different users

Description automatically generatedدر شکل زیر هر دو نمودار نشان می‌دهند که تفاوت زیادی بین مقدار شباهت کاربران یکسان و مقدار شباهت کاربران متفاوت وجود دارد.

۲- شباهت چیدمان روابط: اطلاعات پروفایل کاربر برای همه‌ی افراد در دسترس نیست. برای تجزیه و تحلیل عمیق‌تر مسئله، کاربران یک گروه را بر اساس درجه ارتباطشان به دو دسته تقسیم می‌کنیم: دسته کاربران رهبر و دسته کاربران عادی. کاربران رهبر دارای درجه بالا هستند ولی کاربران عادی درجه کمتری دارند. عملکرد یک گروه معمولا براساس کاربرانی که دوستان بیشتری دارند تعیین می‌شود. ضریب *rich club* (*rc*) نسبت تعداد یال ها به رئوس با درجه بزرگتر از k است. در واقع این ضریب نشان دهنده رابطه بین کاربران رهبر در یک گروه است.

A number and equal sign

Description automatically generated with medium confidenceدر این معادله Nk تعداد رئوس با درجه بیشتر از k است و Ek تعداد یال های بین این رئوس است. از آنجا که تعداد درجه‌های *rc* دو گروه ممکن است تفاوت های زیادی 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 *rc*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 *rc*دو گروه را بر اساس این عبارت می‌سنجیم:

ج) استنباط بر اساس مدل گراف احتمالاتی

مدل گراف 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. برای مثال اگر یک گراف داشته باشیم که رئوس آن، چهار پارامتر هوش، میزان درس خواندن، سختی آزمون و نتیجه کسب شده باشد، با استفاده از این گراف احتمالاتی می‌توان عملکرد دانشجو را پیش‌بینی کرد. در واقع رأس نتیجه کسب شده حاصل برهم نهی ۳ پارامتر دیگر است. [مدل شبکه بیزی](https://github.com/Alisssaaaddd/IEEE-Social-Networks-paper) (*Bayesian*) 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.

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 فراوانی 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. 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.

تحلیل ویژگی‌ها:

شباهت پروفایل بیشترین اهمیت را در نتیجه پایانی دارد که تا 0.607 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، یکسان نیست و این عدد معمولا در شبکه‌های اجتماعی، معنای متفاوتی دارد. (عکس نمودارها)

برای بررسی اینکه آیا یافته‌های ما می‌توانند در برنامه‌های دیگر نیز مفید باشد، مسئله 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. 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 و در نهایت اگر u و v در گروه باشند و ارتباط داشته باشند، مقدار 1 را خروجی می‌دهد.

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.

مقیاس‌پذیری

در بررسی انجام شده، فقط پروفایل کاربران و روابط بین آنها را در نظر گرفتیم. 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 شود. 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 با افزایش منابع محاسباتی به طور قابل توجهی (تقریباً به صورت نمایی) کاهش می‌یابد و همچنین این روش توانایی مقابله با کلان داده‌ها (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.