دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده مهندسی برق و کامپیوتر

پیش‌بینی پیوند در شبکه‌های وزن‌دار

پایان‌نامه برای دریافت درجهٔ کارشناسی ارشد

در رشتهٔ مهندسی کامپیوتر

گرایش هوش مصنوعی و روباتیک

نگارش:

حمید عظیمی

استاد راهنما:

دکتر مسعود اسدپور

شهریور ۱۳۹۴

تقدیم به پدر و مادرم،

که بدانند گرچه با واژه‌ها خوب نیستم،

اما تا همیشه دوستشان دارم و قدرشناسشان هستم.

با تشکر از استاد گرانقدرم، جناب آقای دکتر مسعود اسدپور، که بسیار از ایشان آموخته‌ام و رهنمودهایشان به همراه مهربانی‌شان یاری‌بخش من در مسیرم بوده است.

همچین تشکر می‌کنم از دوستـان عزیز هم‌آزمایشگاهی‌ام که در این دو سال کمک‌های فراوانی به من کردند و لحظه‌های بسیار خوبی را در کنارشان گذراندم، خانم‌ها و آقایان (به ترتیب حروف الفبا!) اکبری دیلمی، دهقانی سانیج، رشیدیان، رفائی افشار قزلباش، شعله، عبدالهی، قوامی، مهدوی لاهیجانی، یدالهی و سایر دوستان خوبم که نامشان از قلم افتاد، که برای تک‌تک‌شان آرزوی موفقیت و شادکامی دارم.

و در آخر با سپاس از دوست بسیار عزیزم خانم نسترن محمودیار که هیچگاه از کوچکترین کمکی دریغ نکرد و صبور و مهربان در کنارم بود و برایش آرزوی بهترین‌ها را دارم.

چکیده

امروزه تحلیل و بررسی شبکه‌های اجتماعی به موضوع مهمی تبدیل شده و توجه پژوهشگران رشته‌های مختلفی را برانگیخته است. در این میان یکی از مسائل مهم موجود، مسئلهٔ پیش‌بینی پیوند است. این مسئله می‌کوشد روابطی که هنوز در یک شبکه شناخته و یا تشکیل نشده‌اند را پیش‌بینی کند. برای حل این مسئله روش‌های بسیاری ارائه شده‌اند. یک دسته از روش‌هایی که برای حل این مسئله وجود دارد، شاخص‌های مبتنی بر شباهت ساختاری هستند که به علت سادگی و کارایی مناسب، محبوبیت زیادی در بین روش‌های پیش‌بینی پیوند دارند. از طرفی در بیشتر پژوهش‌های انجام شده در این زمینه، وزن پیوندها که نشان دهنده قدرت ارتباط است در نظر گرفته نشده است، در حالی که وزن ارتباطات حاوی اطلاعات مفیدی در این راستاست. همچنین می‌توان از اطلاعات ساختاری دیگری مانند انجمن‌های شبکه برای بهبود کارایی پیش‌بینی پیوند استفاده نمود.

هدف اصلی این پژوهش ارائه روشی بر پایهٔ تشخیص انجمن برای پیش‌بینی پیوند در شبکه‌های وزن‌دار است. به منظور تحقق این هدف، با در نظر گرفتن این نکته که احتمال تشکیل ارتباطات درون انجمن‌ها به نسبت بیشتر است، مسئلهٔ پیش‌بینی پیوند درون انجمن‌ها انجام شده‌است. راهکار پیشنهادی دو گام اساسی دارد که با توجه به استفاده یا عدم استفاده از وزن یال‌ها در هر دو گام، به چهار روش گسترش داده می‌شود. به منظور ارزیابی راهکار پیشنهادی، از شبکه‌های مصنوعی LFR استفاده شده‌است که نوعی شبکه پارامتری مقیاس آزاد است. پس از انجام آزمایش روی فضای پارامتری این شبکه‌ها، تحلیلی از کارایی چهار روش پیشنهادی در هر بخش از فضای پارامتری ارائه شده و همچنین شرایطی که منجر به بهبود کارایی روش‌های پیش‌بینی پیوند می‌گردند بررسی می‌شوند.

واژه‌های کلیدی: تحلیل شبکه‌های اجتماعی، پیش‌بینی پیوند وزن‌دار، تشخیص انجمن‌ها، شبکه‌های LFR

فهرست مطالب

فهرست تصاویر

فهرست جداول

فصل ۱: مقدمه

۱-۱- مقدمه

امروزه بسیاری از سامانه‌های اجتماعی، اطلاعاتی و زیستی را می‌توان با استفاده از شبکه‌ها توصیف کرد؛ که در آن، گره‌ها[۱] معرف افراد هستند و یال‌ها[۲] (پیوندهای بین گره‌ها) ارتباط یا تعامل بین گره‌ها را مشخص می‌کنند. شبکه‌های اجتماعی برخط[۳] مانند توییتر[۴]، لینکداین[۵] و...، شبکه‌های پرسش و پاسخ برخط مانند استک اُوِر فِلو[۶]، پاسخ‌های یاهو[۷] و...، شبکه‌های ارتباط بین ژن‌ها یا پروتئین‌ها، زنجیره‌ی غذایی، شبکه‌ی ارتباطی فرودگاه‌های کشور، شبکه‌ی زیرساخت اینترنت، شبکه‌ی برق‌رسانی کشور و... از این دسته‌اند. با توجه به این موضوع، امروزه مطالعه شبکه‌های

[۱] Nodes

[۲] در تمامی این گزارش، کلمه‌های یال ($edge$) و پیوند ($link$) معادل هم هستند.

[۳] Online

[۴] Twitter

[۵] LinkedIn

[۶] StackOverFlow

[۷] Yahoo Answers

پیچیده[۸] تبدیل به یک نقطه توجه مشترک بین شاخه‌های مختلف علم شده است. تلاش‌های بسیاری در راستای فهم سیر تکاملی شبکه‌ها، ارتباط میان همبندی[۹] شبکه و عملکرد آن، و ویژگی‌های شبکه انجام شده‌است. یکی از مسائل علمی مهم مرتبط با تحلیل شبکه‌ها، مسئله‌ایست موسوم به بازیابی اطلاعات[۱۰] که هدف آن، به دست آوردن اطلاعات مفید و مورد نیاز، از یک توده‌ی عظیم داده‌هاست. همچنین می‌توان به این موضوع از منظر پیش‌بینی ارتباطات بین افرادِ شبکه، و به صورت گسترده‌تر مسئله‌ی پیوندکاوی[۱۱]، نگاه کرد که در این میان، مسئله‌ی پیش‌بینی پیوند[۱۲] ، یکی از بنیادی‌ترین مسائل است که سعی دارد تا شانس وجود و یا تشکیل پیوند بین دو گره را، با استفاده از مشاهدات روی پیوندها و ویژگی‌های گره‌ها، تخمین بزند. مسئلهٔ پیش‌بینی پیوند را می‌توان در دو بخش عمده دسته‌بندی کرد: دستهٔ اول پیش‌بینی پیوندهایی است که وجود دارند اما هنوز ناشناخته باقی مانده‌اند، مثل شبکه‌های غذایی، شبکهٔ تعاملی پروتئین-پروتئین و شبکه‌های زیستی[۱۳]؛ دسته‌ی دیگر پیش‌بینی پیوندهایی است که ممکن است در آینده در یک شبکه پویا و در حال تغییر و تکامل، مانند شبکه‌های اجتماعی برخط، ایجاد شوند. در بیشتر تحقیقاتی که تاکنون در این حوزه انجام گرفته، وزن پیوندها لحاظ نشده و همهٔ آن‌ها، هم‌وزن در نظر گرفته شده‌اند، اما در بسیاری از شبکه‌ها، پیوندها دارای وزن هستند و میزان ارتباط بین گره‌ها می‌تواند در پیش‌بینی تاثیرگذار باشد. هدف این پژوهش، بررسی تاثیر مشارکت وزن پیوندها در کیفیت پیش‌بینی است. این پژوهش همچنین می‌کوشد تا از ساختار انجمن‌های شبکه‌ها کمک بگیرد تا بتواند به بهبود کارایی روش‌های پیش‌بینی پیوند کمک کند.

[۸] Complex Networks

[۹] topology

[۱۰] Information Retrival

[۱۱] Link Mining

[۱۲] Link Prediction

[۱۳] Biological Networks

۱-۲- اهمیت موضوع و کاربردهای آن

مسئلهٔ پیش‌بینی پیوند می‌تواند کاربردهای متفاوت و متنوعی در انواع مختلف شبکه‌ها داشته باشد. برای مثال، پیشنهاد کردن کالا به کاربر در وبگاه‌های خرید و فروش برخط نظیر آمازون[۱۴] یا ای‌بِی[۱۵] می‌تواند به عنوان یک مسئلهٔ پیش‌بینی پیوند در شبکه‌های دوبخشی[۱۶] کاربر-کالا در نظر گرفته شود و پیشنهادهای دقیق و مناسب می‌تواند فروش این وبگاه‌ها را به میزان قابل توجهی افزایش دهد. مسئلهٔ پیش‌بینی پیوند همچنین می‌تواند در حوزه‌های دیگری نظیر پیش‌بینی همکاری‌های آینده در شبکه‌های همکاری بین نویسندگان و مؤلفان[۱۷]، تشخیص همکاری‌های زیرزمینی بین تروریست‌ها و... استفاده شود. همچنین از پیش‌بینی پیوند می‌توان برای حل مسائل طبقه‌بندی در گراف‌هایی که به صورت ناقص برچسب‌گذاری شده‌اند[۱۸]، برای تشخیص کارکرد پروتئین‌ها یا تشخیص ایمیل‌های ناخواسته استفاده کرد.

در بسیاری از شبکه‌های زیستی مثل زنجیرهٔ غذایی، شبکهٔ ارتباط میان پروتئین‌ها و شبکه‌های متابولیسمی، برای تشخیص وجود پیوند بین دو گره، می‌بایست آزمایش‌هایی در آزمایشگاه انجام شوند که این آزمایش‌ها معمولاً بسیار هزینه‌بر هستند. دانش ما از این نوع شبکه‌ها بسیار محدود است. برای مثال ٪۸۰ از تعاملات مولکولی در سلول‌های مخمر و ٪۹۹٫۷ در سلول‌های انسان، هنوز ناشناخته‌اند [۱] [۲]. به جای بررسی کردن تمام تعاملات ممکن، پیش‌بینی کردن بر اساس تعاملات شناخته‌شده و سپس تمرکز کردن بر روی پیوندهایی که با احتمال بیشتری وجود دارند، با فرض این که پیش‌بینی ما دقت خوبی داشته باشد، می‌تواند به مقدار قابل توجهی در هزینه‌های آزمایش صرفه‌جویی کند. تحلیل شبکه‌های اجتماعی نیز با موضوع داده‌های گم‌شده مواجه هستند[۳]، که در آن‌جا، الگوریتم‌های پیش‌بینی پیوند نقش مهمی ایفا می‌کنند. به علاوه، داده‌هایی که برای ساخت شبکه‌های زیستی یا

[۱۴] Amazon

[۱۵] eBay

[۱۶] Bipartite Network

[۱۷] Co-authorship Network

[۱۸] Partially Labeled

اجتماعی استفاده می‌شود، ممکن است حاوی اطلاعات نادقیق باشد که این امر باعث می‌شود تا پیوندهای جعلی ایجاد شوند[۴]. مسئلهٔ پیش‌بینی پیوند می‌تواند برای تشخیص این پیوندهای جعلی نیز به کار گرفته شود[۵].

علاوه بر این که الگوریتم‌های پیش‌بینی پیوند می‌توانند در یافتن داده‌های گم‌شده به ما کمک کنند، این الگوریتم‌ها می‌توانند برای پیش‌بینی پیوندهایی که ممکن است در آینده در یک شبکهٔ در حال تغییر و تحول ایجاد شوند نیز مورد استفاده قرار گیرند. برای مثال، در یک شبکهٔ اجتماعی برخط، پیوندهایی که در حال حاضر وجود ندارند اما شانس تشکیل‌شان بسیار بالاست، می‌توانند به عنوان دوستی‌های بالقوه به کاربران[۱۹] آن شبکهٔ اجتماعی پیشنهاد شوند، که همین موضوع می‌تواند کاربران را در یافتن دوست‌های جدید یاری کند و موجب تقویت وفاداری و افزایش میزان استفاده کاربران از شبکهٔ اجتماعی شود. مشابه همین روش می‌تواند برای ارزیابی سازوکار تغییر و تحول یک شبکهٔ مشخص مورد استفاده قرار بگیرد. برای مثال مدل‌های بسیاری برای تغییر و تحول ساختار شبکهٔ جهانی اینترنت ارائه شده است: بعضی از آن‌ها سعی می‌کنند به طور دقیق‌تر توزیع درجه‌ها و نحوهٔ اتصال گره‌ها به هم را بازتولید کنند[۶]، برخی دیگر می‌کوشند ساختارهای k-هسته‌ای را بهتر توصیف کنند[۷] و غیره. از آن‌جایی که ویژگی‌های ساختاری بسیاری برای شبکه‌ها وجود دارد و وزن‌دهی به آن‌ها بسیار سخت است، قضاوت این که کدام مدل (برای مثال کدام سازوکار تغییر و تحول) از بقیه بهتر است، آسان نیست. باید توجه داشت که هر مدل در اصل به یک الگوریتم پیش‌بینی پیوند متناظر می‌شود و بنابراین ما می‌توانیم از معیار دقت پیش‌بینی برای ارزیابی کارایی مدل‌های متفاوت استفاده کنیم.

همان‌طور که مشاهده می‌شود، مسئلهٔ پیش‌بینی پیوند، طیف بسیار وسیعی از کاربرها را شامل می‌شود. با توجه به این موضوع، لزوم یافتن روش‌هایی با کارایی بالا، بدیهی به نظر می‌رسد.

[۱۹] Users

۱-۳- اهداف و دستاوردهای پژوهش

همان‌طور که در بخش قبل خاطر نشان شد، بهبود کارایی روش‌های پیش‌بینی پیوند موضوع بسیار مهمی است که توجه ویژه‌ای را طلب می‌کند. برای افزایش دقت این الگوریتم‌ها، می‌بایست تا جای ممکن اطلاعاتی را که یک شبکه می‌تواند در اختیار ما قرار دهد، استخراج، و از آن‌ها به نحو احسن استفاده کرد.

یکی از اطلاعاتی که معمولاً در شبکه‌ها به آن دسترسی داریم، اطلاعات وزن پیوندهاست. یعنی رابطهٔ بین گره‌ها در این شبکه‌ها فراتر از یک رابطهٔ دودویی[۲۰] است که فقط وجود یا عدم وجود پیوند را نشان دهد، به این معنی که به ما قدرت رابطه بین دو گره را نیز نشان می‌دهد. برای مثال در یک شبکهٔ اجتماعی، میزان تعامل بین دو کاربر می‌تواند وزن رابطهٔ آن‌ها باشد. یا در شبکهٔ فرودگاه‌های کشور، تعداد پرواز بین دو فرودگاه را می‌توان به عنوان وزن پیوند ارتباطی بین آن‌ها در نظر گرفته شود. تلاش‌هایی در زمینهٔ استفاده از وزن پیوندها در پیش‌بینی انجام شده است که در بخش‌های بعدی به آن‌ها اشاره خواهد شد. یکی از اهداف این پژوهش بررسی این نکته است که استفاده از این وزن پیوندها چگونه می‌تواند به ما در بهبود کارایی روش‌ها کمک کند که در انتها به عنوان یکی از دستاوردهای این پژوهش دربارهٔ آن بحث خواهد شد.

یکی دیگر از اطلاعاتی که یک شبکه در اختیار ما قرار می‌دهد، اطلاعات ساختاری آن است و یکی از این اطلاعات ساختاری که می‌تواند از یک شبکه استخراج شود، اطلاعات انجمن‌های[۲۱] آن شبکه است. یکی دیگر از دست‌آوردهای پژوهش حاضر، این است که به وسیلهٔ ترکیب اطلاعات انجمن‌ها با اطلاعات وزن پیوندها، روش جدیدی ارائه می‌دهد که می‌تواند در مواردی که دربارهٔ آن‌ها به تفصیل بحث خواهد شد، کارایی روش‌های پیش‌بینی پیوند را بهبود بخشد.

[۲۰] Binary Relation

[۲۱] Community

۱ - ۴ - ساختار پایان‌نامه

فصل ۲: در این فصل ابتدا مسئلهٔ پیش‌بینی پیوند به طور رسمی معرفی می‌شود و تعاریف، اصول و مبانی نظری این مسئله مورد بررسی قرار می‌گیرد. سپس روش‌های مختلف پیش‌بینی پیوند دسته‌بندی می‌شوند و هر کدام از این روش‌ها به صورت مختصر معرفی می‌شوند. در این بخش روش‌هایی که در این پژوهش مورد استفاده قرار گرفته‌اند با جزئیات بیشتری بررسی خواهند شد. در ادامه به دلیل استفاده از روش‌های تشخیص انجمن در این پژوهش، مروری اجمالی بر این روش‌ها نیز انجام خواهد گرفت و روش مورد نظر معرفی خواهد شد.

فصل ۳: در این فصل ابتدا به بیان مقدمه و پیش‌نیازهای بحث پرداخته می‌شود و سپس روش پیشنهادی این پژوهش معرفی می‌گردد که همان استفاده از اطلاعات انجمن‌ها و پیش‌بینی پیوند در داخل انجمن است. سپس تعریف ریاضی روش بررسی خواهد شد. در ادامه در مورد چگونگی تاثیر وزن یال‌ها در روش پیشنهادی بحث خواهد شد. سپس یک رویکرد عملی متفاوت برای محاسبهٔ تقریبی شاخص‌های پیشنهادی ارائه شده و در مورد نقاط قوت و ضعف آن صحبت خواهد شد. در پایان نیز جمع‌بندی کوتاهی از این فصل ارائه می‌شود.

فصل ۴: در این فصل ابتدا مجموعه داده‌های مورد استفاده در این پژوهش معرفی می‌شوند. سپس معیارهای ارزیابی روش‌های پیش‌بینی پیوند به طور کامل مورد بحث قرار می‌گیرند. در نحوه انجام آزمایش‌ها توضیح داده می‌شود و نتایج به دست آمده از آن‌ها در قالب نمودارها و جداول ارائه شده و با توجه به معیارهای مختلف ارزیابی تشریح و تفسیر، و با هم مقایسه می‌شوند.

فصل ۵: در این فصل نیز که فصل پایانی این پژوهش است، یک جمع‌بندی از مطالب ارائه‌شده در این پژوهش بیان می‌شود. و در نهایت کارهای آینده‌ای که در راستای این پژوهش می‌توانند مورد توجه قرار گیرند، بحث و بررسی خواهند شد.

فصل ۲: مروری بر منابع

۲-۱- مقدمه

در این بخش، ابتدا به تعریف مسئلهٔ پیش‌بینی پیوند پرداخته می‌شود و این مسئله به صورتی رسمی و ریاضی مدل خواهد شد. نخست به تعریف عمومی مسئله پرداخته می‌شود و سپس به حالت وزن‌دار که در این پژوهش مد نظر است اشاره خواهد شد. در ادامه روش‌های مختلف پیش‌بینی پیوند بررسی می‌شوند و دربارهٔ هر کدام توضیح مختصری بیان خواهد شد . دسته روش‌هایی که پایهٔ راهکار ارائه‌شده در این پژوهش است یعنی روش‌های مبتنی بر معیارهای شباهت محلی با جزییات بیشتری بررسی می‌شوند. در بخش بعد به توضیح مختصری در باب روش‌های تشخیص انجمن‌ها پرداخته خواهد شد؛ چرا که در روش پیشنهادی که در بخش‌های آیندهٔ این پژوهش ارائه می‌شوند، از روش‌های تشخیص انجمن‌ها بهره برده شده‌است.

۲-۲- تعاریف، اصول و مبانی نظری

یک گراف ساده بدون جهت $G(V,E)$ را در نظر بگیرید، که در آن V مجموعهٔ گره‌ها و E مجموعهٔ یال‌های بین گره‌هاست. طبق تعریف گراف ساده، یال‌های چندگانه[۱] و حلقه‌ها[۲] در این گراف مجاز نیستند. مجموعهٔ تمام یال‌های ممکن بین تمام گره‌ها را با U نشان می‌دهیم که تعداد این یال‌ها $\frac{|V|\times(|V|-1)}{2}$ است. نماد $|V|$ به معنی تعداد اعضای مجموعهٔ V است که در واقع تعداد یال‌ها را نشان می‌دهد. در نتیجه مجموعهٔ یال‌های ناموجود در گراف G برابر است با $U-E$. ما فرض می‌کنیم که در این مجموعه تعدادی پیوند گم‌شده وجود دارند (پیوندهایی که دیده نشده‌اند و یا ممکن است در آینده به وجود بیایند) و هدف ما این است که این پیوندهای گم‌شده را پیدا کنیم و پیش‌بینی کنیم.

در حالت کلی ما نمی‌دانیم کدام‌یک از پیوندها دیده نشده‌اند یا ممکن است در آینده به وجود بیایند، چون در غیر این صورت دیگر نیازی به پیش‌بینی نداشتیم. بنابراین برای آزمودن دقت الگوریتم‌ها، مجموعهٔ پیوندهای گراف G یعنی E را به صورت تصادفی به دو بخش افراز می‌کنیم: بخش اول دادهٔ آموزش که با E' نمایش می‌دهیم و به عنوان دادهٔ شناخته از آن استفاده می‌کنیم؛ و بخش دوم که با E'' نمایش داده می‌شود و دادهٔ آزمون ماست و برای آزمودن دقت الگوریتم‌ها استفاده می‌شود و هیچ اطلاعاتی از آن نمی‌تواند در فرآیند پیش‌بینی استفاده شود. طبق تعریف افراز واضح است که $E' \cup E'' = E$ و $E' \cap E'' = E$ برقرارند. مزیت این افراز تصادفی این است که نسبت تقسیم، به تعداد تکرارها وابسته نیست. اما با این روش، بعضی پیوندها ممکن است هیچ‌گاه در مجموعهٔ آزمون قرار نگیرند، و از طرف دیگر بعضی بیش از یک بار در مجموعهٔ آزمون قرار بگیرند که همین باعث سوگیری[۳] آماری می‌شود. برای بر طرف کردن این محدودیت می‌توان از روش ارزیابی متقاطع K-قسمتی[۴] استفاده کرد که در آن، مجموعهٔ پیوندهای مشاهده شده (E) به صورت تصادفی به K زیرمجموعه افراز می‌شوند. هر بار یکی از این زیرمجموعه‌ها به عنوان مجموعهٔ آزمون انتخاب می‌شود و اجتماع بقیهٔ $K-1$ زیرمجموعه به عنوان مجموعهٔ آزمون استفاده می‌شود.

[۱] Multiple Edge

[۲] Self Loops

[۳] Bias

[۴] K-fold cross-validation

این فرآیند K بار تکرار می‌شود و بنابراین هر زیرمجموعه دقیقاً یک بار به عنوان مجموعهٔ آزمون انتخاب شود. با این کار، تمام پیوندها هم برای آموزش و هم برای اعتبارسنجی مورد استفاده قرار می‌گیرند و هر پیوند دقیقاً یک بار برای پیش‌بینی به کار می‌رود. به وضوح هر چقدر K بیشتر باشد، سوگیری آماری کمتری خواهیم داشت، اما از طرف دیگر هزینهٔ محاسباتی بیشتری را می‌بایست متقبل شویم. بعضی شواهد تجربی پیشنهاد می‌کنند که استفاده از روش ارزیابی متقاطع ۱۰-قسمتی، توازن[۵] بسیار خوبی بین هزینه و کارایی ایجاد می‌کند[۸] و [۹]. در این پژوهش نیز از همین روش استفاده خواهد شد.

همان‌طور که پیش‌تر گفته شد، هدف ما بررسی تاثیر مشارکت وزن پیوندها در کیفیت پیش‌بینی پیوند است، بنابراین گراف ورودی ما یک گراف وزن‌دار خواهد بود. گسترش تعریف این مسئله به مسئله پیش‌بینی پیوند وزن‌دار بسیار ساده است. تنها فرض اضافه شده این است که به هر پیوند یک عدد مثبت به عنوان وزن آن پیوند اضافه شده‌است. در ماتریس مجاورت گراف‌های بدون وزن، اعداد ۱ و ۰ به ترتیب به معنی وجود و عدم وجود پیوند هستند، در حالی که در ماتریس مجاورت گراف‌های وزن‌دار، هر عدد وزن پیوند متناظر را مشخص می‌کند.

۲-۳- مروری بر ادبیات پیش‌بینی پیوند

مسئلهٔ پیش‌بینی پیوند یک چالش قدیمی در دانش اطلاعات مدرن است و الگوریتم‌های بسیاری در این زمینه ارائه شده‌اند. این الگوریتم‌ها طیف وسیعی را شامل می‌شوند که از روش‌های برپایهٔ شباهت گره‌ها گرفته تا روش‌های بر پایهٔ زنجیره‌های مارکوف[۶] و مدل‌های آماری، گسترده شده‌اند. دسته‌بندی‌های مختلفی توسط افراد مختلف از این روش‌ها ارائه شده‌است. برای مثال لو[۷] و ژو[۸][۱۰] در سال ۲۰۱۱ این روش‌ها را به سه دستهٔ کلی زیر تقسیم کرده‌اند:

[۵] Trade-off

[۶] Markov chains

[۷] Lu

[۸] Zhou

۱. الگوریتم‌های بر پایهٔ شباهت[۹]

۲. روش‌های بیشینه همانندی [۱۰]

۳. مدل‌های احتمالاتی[۱۱]

یک دسته‌بندی دیگر از روش‌ها توسط ونگ[۱۲] و همکاران[۱۱] در سال ۲۰۱۵ ارائه شد که روش‌ها را به چهار دستهٔ کلی زیر تقسیم‌بندی می‌کند:

۱. معیارهای بر پایهٔ گره[۱۳]

۲. معیارهای بر پایهٔ هم‌بندی [۱۴]

۳. معیارهای بر پایهٔ نظریه اجتماعی[۱۵]

۴. روش‌های بر پایهٔ یادگیری[۱۶]

در ادامه به توضیح مختصری دربارهٔ هر کدام از دسته روش‌ها پرداخته خواهد شد و روش‌هایی که مد نظر این پژوهش هستند به تفصیل مورد بررسی قرار خواهند گرفت.

۲-۳-۱ معیارهای بر پایهٔ گره

محاسبهٔ شباهت بین یک جفت گره راه حلی بدیهی و شهودی برای پیش‌بینی پیوند است. این معیار بر پایهٔ ایده‌ای ساده استوار است: «جفت گره‌ای شانس وجود پیوند بین آن‌ها بیشتر است که بیشتر شبیه به هم هستند و بالعکس»[۱۱].

[۹] Similarity-based Algorithms
[۱۰] Maximum Likelihood
[۱۱] Probabilistic Models
[۱۲] Wang
[۱۳] Node-based Metrics
[۱۴] Topology-based Metrics
[۱۵] Social Theory based Metrics
[۱۶] Learning-based Methods

این ایده بر اساس این واقعیت است که کاربران به ایجاد روابط با افرادی که در آموزش، مذهب، علایق، مکان و... مشابه آن‌ها می‌باشند، تمایل دارند. تشابه این‌گونه اندازه‌گیری می‌شود که به هر جفت غیرمتصل از گره‌ها مثل (x, y) نمرهٔ مفهوم شباهت بین آن دو تخصیص می‌شود. واضح است که نمرهٔ بالا نشان‌دهندهٔ احتمال بیشتر ایجاد پیوند بین x و y در آینده خواهد بود، و به طور مشخص نمره کم نیز نشان می دهد که به احتمال زیاد دو گره x و y به یکدیگر متصل نخواهند شد. بنابراین، با استفاده از رتبهٔ نمرات تشابه بین گره‌ها، می‌توان تشکیل یا عدم تشکیل پیوندهایی در آینده و یا پیوند نهان در شبکه فعلی را پیش‌بینی کرد. در یک شبکه اجتماعی عملی ، یک گره معمولا دارای برخی ویژگی ها از قبیل مشخصات کاربری در شبکه‌های اجتماعی برخط، سابقه انتشار[۱۷] در شبکه‌های اجتماعی دانشگاهی و... است. این اطلاعات می‌تواند به طور مستقیم برای محاسبهٔ شباهت بین دو گره استفاده شود. از آن‌جا که در بیشتر موارد، مقادیر ویژگی‌های گره‌ها به شکل متنی هستند، معیارهای شباهتی که معمولا مورد استفاده قرار می‌گیرند بر پایهٔ متن[۱۸] و بر پایهٔ رشته[۱۹] هستند. در مقالهٔ [۱۲] یک مدل درختی طبقه‌بندی چندگانه تعریف شده که به مطالعهٔ کلمات کلیدی[۲۰] پروفایل کاربر می‌پردازد، سپس فاصلهٔ بین کلمات کلیدی را تعیین می‌کنند تا شباهت بین هر جفت از کاربران مشخص شود. در نتیجه، معیارهای بر پایهٔ گره عمدتا از ویژگی‌ها فردی و فعالیت‌های کاربران استفاده می‌کند، که می‌تواند منعکس‌کنندهٔ علایق شخصی و رفتارهای اجتماعی آنها برای محاسبه شباهت بین جفت گره‌ها باشد. بنابراین، معیارهای بر پایهٔ گره در پیش‌بینی پیوند مفید هستند به شرطی که بتوانیم ویژگی‌ها فردی و فعالیت‌های کاربران را در شبکه‌های اجتماعی به دست آوریم[۱۳].

[۱۷] Publication Record

[۱۸] Text-based

[۱۹] String-based

[۲۰] Keywords

۲-۳-۲ روش‌های بر پایهٔ شباهت

ساده‌ترین چارچوب روش‌های پیش‌بینی پیوند، الگوریتم‌های بر پایهٔ شباهت هستند. در بعضی منابع به این روش‌ها روش‌های بر پایهٔ همبندی[۲۱] یا روش‌های بر پایهٔ شباهت ساختاری[۲۲] نیز گفته می‌شود.

با وجود سادگی آن‌ها، مطالعه بر روی الگوریتم‌های برپایهٔ شباهت خود موضوع مهمی است. در حقیقت، تعریف شباهت گره یک چالش کوچک اما بااهمیت است. شاخص تشابه می‌تواند بسیار آسان یا بسیار پیچیده باشد. هر شاخص ممکن است برای برخی شبکه‌ها به خوبی پاسخگو باشد و اما در عین حال برای بعضی شبکه‌های دیگر با شکست مواجه شود. همان‌طور که در بخش پیش گفته شد، شباهت گره می‌تواند با استفاده از ویژگی‌های اساسی گره‌ها تعریف شود. دو گره مشابه در نظر گرفته می‌شوند اگر که ویژگی‌های مشترک زیادی با یکدیگر داشته باشند. اما از آن‌جایی که ویژگی‌های گره‌ها عموماً در دسترس نیستند، در نتیجه معمولاً بر گروه دیگری از شاخص‌های شباهت با نام شباهت ساختاری متمرکز می‌شوند که فقط بر پایهٔ ساختار شبکه استوار است.

شاخص‌های شباهت ساختاری را می‌توان به روش‌های متعددی طبقه‌بندی کرد، از جمله طبقه‌بندی‌ها می‌توان به محلی[۲۳] در مقابل سراسری[۲۴]، بدون پارامتر در قیاس با وابسته به پارامتر، وابسته به گره در مقابل وابسته به مسیر و غیره اشاره کرد.

شاخص‌های شباهت محلی

این روش‌ها، زیرشاخه‌ای از روش‌های بر پایهٔ شباهت هستند که از اطلاعات محلی، یعنی اطلاعات خود گره‌ها و همسایگان آن‌ها استفاده می‌کنند. این روش‌ها به دلیل سادگی مفهومی و محاسباتی و همچنین کارایی مناسب، از

[۲۱] Topology-based
[۲۲] Structural Similarity
[۲۳] Local
[۲۴] Global

محبوبیت بسیاری برخوردارند. پژوهش حاضر نیز از همین دسته از روش‌ها استفاده خواهد کرد. در زیر به معرفی این روش‌ها و توضیح جزییات آن‌ها پرداخته می‌شود.

همسایگان مشترک (CN)[۲۵]: شاخص همسایگان مشترک به دلیل سادگی یکی از گسترده‌ترین شاخص‌های مورد استفاده در مسائل پیش‌بینی پیوند است. برای هر دو گره x و y، این شاخص معرف تعداد گره‌هایی است که با هر دو گره x و y ارتباط مستقیمی داشته باشد و در واقع همسایهٔ مشترک هر دو گره باشند. بنابراین، دو گره x و y چنانچه همسایه‌های مشترک بسیاری داشته باشند با احتمال خوبی با یکدیگر نیز پیوند دارند. ساده‌ترین اندازه‌گیری این همپوشانی همسایه‌ها، شمارش مستقیم است. برای یک گره با نام x،مجموعه‌ای از همسایه‌های x با $\Gamma(x)$ نشان داده می‌شود:

$$S_{xy}^{CN} = |\Gamma x \cap \Gamma y| \qquad (۱-۲)$$

که $|Q|$ تعداد اعضای مجموعهٔ Q است. بدیهی است که $S_{xy} = (A^2)_{xy}$، که در آن A ماتریس مجاورت[۲۶] است و چنانچه x و y به طور مستقیم با یکدیگر در ارتباط باشند و پیوندی بین آن‌ها باشد $A_{xy} = 1$ و در غیر اینصورت $A_{xy} = 0$ می‌باشد.توجه شود که همچنین $(A^2)_{xy}$ بیانگر تعداد مسیرهای مختلف به طول ۲ است که x و y را به هم متصل می‌کنند. نیومن[۲۷][۱۴] این کمیت را در مطالعات خود در زمینهٔ شبکه‌های همکاری مورد بررسی قرار داد و نشان داد همبستگی مثبتی بین تعداد همسایگان مشترک و احتمال این که دو محقق در آینده همکاری داشته باشند، وجود دارد. کاسینتز[۲۸] و واتس[۲۹][۱۵] یک شبکه‌های اجتماعی با مقیاس بزرگ را تجزیه و تحلیل کردند که نشان می‌دهد دو دانشجو که دارای دوستان مشترک بسیاری هستند، دوست شدن آن‌ها در آینده از احتمال خوبی برخوردار است. از آن‌جا که در این پژوهش هدف استفاده از وزن‌یال‌هاست، نیاز

[۲۵] Common Neighbors

[۲۶] Adjacency Matrix

[۲۷] Newman

[۲۸] Kossinets

[۲۹] Watts

داریم که از گسترش وزن‌دار شاخص‌ها استفاده کنیم. موراتا[۳۰] و موریاسو[۳۱] در سال ۲۰۰۷ گسترش وزن‌داری از سه شاخص شباهت ارائه دادند [۱۶]. اولین شاخص، شاخص همسایگان مشترک است که رابطهٔ گسترش وزن‌دار آن به صورت زیر است. دو شاخص دیگر در ادامه بررسی خواهند شد.

$$S_{xy}^{WCN} = \frac{1}{2} \sum_{z \in \Gamma(x) \cup \Gamma(y)} w(x,z) + w(z,y) \qquad (۲-۲)$$

از آن‌جا که شاخص همسایگان مشترک نرمال‌شده[۳۲] نیست، معمولا شباهت نسبی بین جفت گره‌ها را نشان می‌دهد. بنابراین، برخی از معیارهای دیگر بر پایهٔ همسایه‌ها، بررسی می‌کنند که چگونه می‌توان این معیار را به شکل منطقی نرمال کرد.

شاخص سالتون[۳۳]: این شاخص یک شاخص کوسینوسی برای محاسبهٔ شباهت بین دو گره x و y است و به شکل زیر تعریف می‌شود [۱۷]:

$$S_{xy}^{Salton} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\sqrt{k_x \times k_y}} \qquad (۲-۳)$$

در این‌جا k_x نشان‌دهندهٔ درجهٔ گره x است. شاخص سالتون در بعضی مواقع تشابه کسینوسی نیز نامیده می‌شود.

شاخص جاکارد[۳۴]: این شاخص که اندازه همسایگان مشترک را نرمال می‌کند توسط جاکارد بالغ بر ۱۰۰ سال پیش ارائه شد. این شاخص فرض می‌کند که شباهت بیشتر، برای زوج گره‌هایی است که نسبت بالاتری از مجموع

[۳۰] Murata
[۳۱] Moriyasu
[۳۲] Normalized
[۳۳] Salton Index
[۳۴] Jaccard Index

همسایگانشان بین آن‌ها مشترک است و به شکل زیر تعریف می‌شود:

$$S_{xy}^{Jaccard} = \frac{|\Gamma(x) \cap \Gamma(y)|}{|\Gamma(x) \cup \Gamma(y)|} \quad (۴-۲)$$

شاخص سورنسن[۳۵]: این شاخص علاوه بر توجه به اندازهٔ همسایه‌های مشترک، همچنین بیان می‌کند که گره‌هایی با مجموع درجهٔ پایین‌تر شانس ایجاد پیوند بالاتری دارند. این شاخص عمدتاً برای داده‌های مربوط به محیط زیست مورد استفاده قرار می‌گرفته‌است و به صورت زیر تعریف می‌شود [۱۸]:

$$S_{xy}^{Sorensen} = \frac{2|\Gamma(x) \cap \Gamma(y)|}{k_x + k_y} \quad (۵-۲)$$

شاخص HP[۳۶]: این شاخص همپوشانی همبندانه[۳۷] بین دو گره را محاسبه می‌کند و برای استفاده در شبکه‌های زیستی پیشنهاد شده است. این شاخص به صورت زیر تعریف می‌شود:

$$S_{xy}^{HPI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\min\{k_x, k_y\}} \quad (۶-۲)$$

بر اساس این تعریف، پیوندهای مجاور به هاب‌ها محتمل‌تر هستند که نمرات بالاتری به آن‌ها تخصیص یابد به این دلیل که درجهٔ کوچک‌تر، مخرج را تعیین می‌کند. به بیانی دیگر، ارزش این شاخص توسط گره‌های با درجه کمتر تعیین می‌شود [۱۹].

شاخص HD[۳۸]: ژو و همکاران پیشنهاد یک شاخص مشابه HPI را مطرح کردند [۲۰]، اما ارزش را گره‌های با درجهٔ بالاتر

[۳۵] Sorensen Index

[۳۶] Hub Promoted Index

[۳۷] Topological

[۳۸] Hub Depressed Index

تعیین می‌کنند. این شاخص به صورت زیر تعریف می‌شود:

$$S_{xy}^{HDI} = \frac{|\Gamma(x) \cap \Gamma(y)|}{\max\{k_x, k_y\}} \quad (۷-۲)$$

شاخص LHN[۳۹]: شاخص لایت-هولم-نیومن یا LHN، شباهت بیشتر را به زوج گره‌هایی که همسایگان بیشتری (نه در قیاس با بیشینه همسایگان محتمل بلکه) در مقایسه با تعداد همسایگان مورد انتظار دارند، تخصیص می‌دهد [۲۱] و به صورت زیر تعریف می‌شود:

$$S_{xy}^{LHN1} = \frac{|\Gamma(x) \cap \Gamma(y)|}{k_x \times k_y} \quad (۸-۲)$$

همان‌طور که مشاهده می‌شود مخرج کسر بالا ($k_x \times k_y$) متناسب با تعداد همسایگان مورد انتظار مشترک گره‌های x و y است.

شاخص وابسته به پارامتر (PD)[۴۰]: به منظور بهبود دقت برای پیش‌بینی هر دو دستهٔ پیوندهای محبوب و غیرمحبوب، ژو و همکاران معیار PD را به شرح زیر پیشنهاد کردند [۲۲]. در این‌جا λ یک پارامتر آزاد است. زمانی که $\lambda = 0$ باشد، معیار PD همان معیار همسایگان مشترک (CN) است و اگر $\lambda = 0.5$ و یا $\lambda = 1$ باشد به ترتیب معادل معیارهای سالتون و LHN می‌شود.

$$S_{xy}^{PD} = \frac{|\Gamma(x) \cap \Gamma(y)|}{(|k_x| \times |k_y|)^{\lambda}} \quad (۹-۲)$$

شاخص وابستگی ترجیحی (PA)[۴۱]: این شاخص نشان می‌دهد که پیوندهای جدید، بیشتر احتمال دارد به گره‌هایی با درجه بالاتر متصل شوند. این شاخص می‌تواند به منظور ایجاد تغییر و تحول در شبکه‌های مقیاس آزاد به کار

[۳۹] Leicht-Holme-Newman

[۴۰] Parameter Dependent

[۴۱] Prefrential Attachment

گرفته شود، که در آن احتمال این که یک پیوند جدید به گره x متصل شود متناسب با k_x است. همچنین می‌توان از راهکار مشابهی در شبکه‌های مقیاس آزاد بدون رشد استفاده کرد که در آن در هر مرحله، یک پیوند قدیمی حذف شده و یک پیوند جدید تولید می‌شود و احتمال آنکه پیوند جدید دو گره x و y را به یکدیگر متصل کند متناسب با $k_x \times k_y$ است [۱۴].

$$S_{xy}^{PA} = k_x \times k_y \qquad (۲-۱۰)$$

باید توجه داشت که این معیار به اطلاعات همسایگان هر گره نیازی ندارد، در نتیجه از حداقل پیچیدگی محاسباتی برخوردار است. این شاخص یکی دیگر از شاخص‌هاییست که توسط موراتا و موریاسو به حالت وزن‌دار گسترش داده شد. گسترش وزن‌دار این معیار نیز به صورت زیر است:

$$S_{xy}^{WPA} = s(x) \times s(y) \qquad (۲-۱۱)$$

که در آن، $s(x)$ قدرت [۴۲] گره x را مشخص می‌کند که عبارتست از مجموع وزن یال‌های متصل به گره x و یا به صورت ریاضی $s(x) = \sum_{i \in \Gamma(x)} w(x,i)$.

شاخص آدامیک /ادار (AA) [۴۳]: این شاخص با شمارش سادهٔ همسایه‌های مشترک با اختصاص وزن بیشتر به همسایگانی که خود آن‌ها دارای همسایگان کمتری هستند، تعریف می‌شود. این معیار توسط آدامیک [۴۴] و ادار [۴۵] در ابتدا برای محاسبهٔ شباهت بین دو صفحه وب پیشنهاد شد [۲۳]، که پس از آن به طور گسترده‌ای در شبکه‌های

[۴۲] strength

[۴۳] Adamic/Adar

[۴۴] Adamic

[۴۵] Adar

اجتماعی مورد استفاده قرار گرفت. رابطهٔ محاسبهٔ معیار AA به شکل زیر تعریف شده‌است:

$$S_{xy}^{AA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{\log k_z}. \qquad (۱۲-۲)$$

سومین شاخصی که گسترش وزن‌دار آن توسط موراتا و موریاسو معرفی شد، شاخص AA است که به صورت زیر به نسخه وزن‌دار گسترش داده می‌شود:

$$S_{xy}^{WAA} = \frac{1}{2} \sum_{z \in \Gamma(x) \cup \Gamma(y)} \frac{w(x,z) + w(z,y)}{log(1 + s(z))} \qquad (۱۳-۲)$$

که جمع کردن عدد یک در مخرج به دلیل پرهیز از منفی شدن امتیازها برای مواقعی‌ست که وزن‌ها از یک کوچک‌ترند.

شاخص تخصیص منابع (RA)[۴۶]: معیار تخصیص منابع یا به اختصار RA توسط ژو و همکاران ارائه شده‌است [۲۰]. این معیار از فرآیندهای فیزیکی تخصیص منابع در شبکه‌های پیچیده الهام گرفته شده‌است. یک جفت گره را در نظر بگیرید، گره x و y، که به صورت مستقیم به یکدیگر مرتبط نیستند. گره x می‌تواند تعدادی منبع را به گره y به وسیلهٔ همسایه‌های مشترک آن دو که نقش انتقال‌دهنده را ایفا می‌کنند، ارسال کند. در ساده‌ترین حالت، فرض می‌کنیم که هر انتقال‌دهنده دارای یک واحد از منابع است و به یک اندازه آن را بین تمام همسایگان خود توزیع می‌کند. شباهت بین x و y می‌تواند به عنوان مقدار منابعی که y از x دریافت می‌کند تعریف شود:

$$S_{xy}^{RA} = \sum_{z \in \Gamma(x) \cap \Gamma(y)} \frac{1}{k_z}. \qquad (۱۴-۲)$$

واضح است که این اندازه‌گیری متقارن است یعنی $S_{yx} = S_{xy}$. معیار RA مشابه AA است، هر دو معیار سهم تاثیر همسایگان مشترک با درجهٔ بالا را کاهش می‌دهند. با این وجود معیار RA به نسبت AA، سهم همسایگان

[۴۶] Resource Allocation

مشترک درجهٔ بالا را شدیدتر می‌کاهد. معیار AA فرمی اینگونه دارد $(\log k_z)^{-1}$ در حالی که معیار RA فرمی به صورت k_z^{-1} دارد. بنابراین، RA و AA برای شبکه‌هایی که میانگین درجه گره‌های آن‌ها کم است، نتایج پیش‌بینی بسیار نزدیکی دارند در صورتی که RA برای شبکه‌هایی با میانگین درجهٔ بالا بهتر عمل می‌کند. علاوه بر این، RA و AA نه تنها از همسایگان مستقیم استفاده می‌کنند بلکه همسایگان همسایگان را نیز در نظر می‌گیرد و همین امر این دو شاخص را از سایر شاخص‌ها متمایز می‌کند. این شاخص نیز توسط لو و ژو در سال ۲۰۱۰ با پیروی از نحوهٔ گسترشی که موراتا و موریاسو ارائه کرده بودند، به حالت وزن‌دار گسترش یافت [۲۴] که گسترش‌یافتهٔ آن به شکل زیر است:

$$S_{xy}^{WRA}(x,y) = \sum_{z \in \Gamma(x) \cup \Gamma(y)} \frac{w(x,z) + w(z,y)}{s(z)} \qquad (۱۵-۲)$$

به این دلیل که همسایگان می‌توانند به طور غیرمستقیم رفتار اجتماعی کاربران را منعکس کنند و به صورت مستقیم بر انتخاب اجتماعی کاربران تاثیرگذار هستند، بسیاری از روش‌های پیش‌بینی پیوند بر پایهٔ همسایه‌ها استوار است.

شاخص‌های بر پایهٔ مسیر

یکی دیگر از زیرمجموعه‌های روش‌های بر پایهٔ شباهت، شاخص‌های بر پایهٔ مسیر هستند که بر خلاف دستهٔ پیش، فقط از اطلاعات محلی بهره نمی‌گیرند، بلکه علاوه بر آن، اطلاعات مسیرهای بین دو گره را نیز مورد استفاده قرار می‌دهند. در زیر به معرفی تعدادی از این شاخص‌ها پرداخته می‌شود.

مسیر محلی[۴۷]: شاخص مسیر محلی یا به اختصار LP از اطلاعات مسیرهای محلی با طول ۲ و ۳ استفاده می‌کند. برخلاف شاخص‌هایی که تنها اطلاعات نزدیک‌ترین همسایه‌ها را به کار می‌برند، این شاخص برخی از اطلاعات اضافی همسایه‌ها به فاصله‌ای با طول ۳ تا گره فعلی را مورد استفاده قرار می‌دهد. بدیهی است که مسیرهای با

[۴۷] Local Path

طول ۲ از مسیرهای با طول ۳ مناسب‌تر هستند [۲۵]. بنابراین یک ضریب تنظیم[۴۸] α برای کاهش اثر مسیرهای با طول ۳ وجود دارد. مقدار α باید عددی کوچک و نزدیک به صفر باشد. این شاخص به شکل زیر تعریف شده‌است. در این‌جا، A^2 و A^3 نشان‌دهندهٔ ماتریس مجاورت گره‌هایی است که مسیر با طول ۲ و ۳ را دارا هستند. در نتیجه، LP نیز ماتریس مجاورتی است که جفت گره‌ها با فاصله‌های به طول ۲ و ۳ را توصیف می‌کند [۲۶].

$$LP = A^2 + \alpha A^3 \qquad (۱۶-۲)$$

کتز[۴۹]: شاخص کتز بر پایهٔ ترکیبی از تمام مسیرهای بین دو گره است. در این روش از تمام مسیرهای بین دو گره استفاده می‌شود. شاخص کتز تعداد و طول تمام مسیرهای موجود بین دو بین گره x و y را به منظور پیش‌بینی پیوند به کار می‌گیرد [۲۷]. میرا شدن نمایی تاثیر مسیرها توسط طول آن‌ها می‌تواند وزن بیشتری را به مسیرهای کوتاه‌تر بدهد. این اندازه‌گیری به شرح زیر است، که در آن $path^l_{xy}$ که مجموعه‌ای از تمام مسیرها از x به y است با طول l است و $\beta > 0$. مقدار بسیار کوچک β باعث می‌شود که کتز بسیار شبیه به شاخص CN شود، چرا که در این صورت، مسیرهای طولانی در شباهت نهایی، تاثیر بسیار کمی دارند.

$$Katz(x,y) = \sum_{l=1}^{\infty} \beta^l . |path^l_{x,y}| = \beta A + \beta^2 A2 + \beta^3 A3 + ... \qquad (۱۷-۲)$$

پیوند دوستان[۵۰]: شاخص پیوند دوستان یا به اختصار FL محاسبهٔ شباهت بین گره x و y با پیمودن تمام مسیرهای با طول محدود به یک کران مشخص است. این شاخص می‌تواند (به دلیل در نظر گرفتن کران) پیش‌بینی پیوند دقیق‌تر و سریع‌تری را ارائه کند. پیوند دوستان فرض می‌کند که افراد در یک شبکهٔ اجتماعی می‌توانند تمام مسیرهای بین خود را به نسبت طول مسیر به کار گیرند [۲۸]. شباهت بین گره x و y به عنوان تعداد مسیرهای

[۴۸] Adjustment Factor

[۴۹] Katz

[۵۰] FriendLink

با طول مختلف l از x تا y به شکل زیر تعریف می‌شود :

$$FL(x,y) = \sum_{i=1}^{l} \frac{1}{i-1} \cdot \frac{|paths_{x,y}^{i}|}{\prod_{j=2}^{i}(n-j)} \qquad (۱۸-۲)$$

که در آن n تعداد گره‌های موجود در شبکه است، و l طول مسیر بین x و y است (به استثنای مسیر با دور)، و $paths_{x,y}^{i}$ مجموعه‌ای از تمام مسیرهای از گره x تا y با طول i است. با این حال، این به آن معنا نیست که lهای بالاتر دقت عمل بیشتری را موجب می‌شوند. در واقع، با زیاد شدن بیش از حد l، دقت به مرور افت خواهد کرد.

شاخص‌های بر پایه ولگشت

این روش‌ها می‌کوشند تا شباهت بین دو گره را با استفاده از ولگشت[۵۱] (یا قدم زدن تصادفی) به دست آورند؛ به این صورت که از احتمال رفتن از یک گره به همسایه‌های آن در ولگشت استفاده می‌کنند و به این صورت معیاری شبیه فاصله بین گره‌ها به دست می‌آورند. در زیر به تعدادی از مهمترین شاخص‌های این دسته پرداخته خواهد شد.

سیم‌رنک[۵۲]: شاخص سیم‌رنک، با این فرض تعریف شده که دو گره مشابه هستند، اگر به گره‌های مشابه متصل باشند. γ پارامتری است که کنترل می‌کند با چه سرعتی وزن گره‌های متصل به هم هنگامی که از گره اصلی دور می‌شوند، کاهش یابد [۲۹].

$$simRank(x,y) = \begin{cases} 1 & \text{if} \quad x = y, \\ \gamma . \frac{\sum_{a \in \Gamma(x)} \sum_{b \in \Gamma(y)} simRank(a,b)}{|\Gamma(x)| . |\Gamma(y)|} & \text{if} \quad otherwise. \end{cases} \qquad (۱۹-۲)$$

سیم‌رنک را می توان با استفاده از مدل پیمایشگر جفت تصادفی[۵۳] توضیح داد: $simRank(x,y)$ تعیین می‌کند

[۵۱] Random Walk

[۵۲] SimRank

[۵۳] random surfer-pairs model

که انتظار می‌رود دو پیمایشگر تصادفی که حرکت خود را از دو گره x و y آغاز کرده‌اند چقدر زود در یک گره یکدیگر را ملاقات کنند. پیچیدگی زمانی سیم‌رنک $O(n^4)$ است که n نمایانگر تعداد گرهٔ شبکه است [۳۰]. به دلیل همین پیچیدگی زمانی بالا، این شاخص برای استفاده در شبکه‌های با مقیاس بزرگ مناسب نیست.

رتبه‌صفحه ریشه‌دار[۵۴]: این شاخص نسخه اصلاح‌شدهٔ الگوریتم رتبه‌صفحه[۵۵] است که در موتورهای جستجو به منظور رتبه‌بندی صفحات وب به کار می‌رود [۳۱]. در این شاخص به هر گره، احتمالی تخصیص داده می‌شود که نشان‌دهندهٔ احتمال رسیدن یک ولگشت به آن گره است. عامل ϵ ضریبی است که مشخص می‌کند که چقدر امکان دارد ولگشت به جای برگشت به گره مبدا، گره همسایه را ملاقات کند.

$$RPR = (1-\epsilon)(I - \epsilon D^{-1}A^{-1})^{-1}, \quad D_{i,i} = \sum_j A_{i,j} \qquad (۲۰-۲)$$

که I ماتریس واحد و A ماتریس مجاورت گراف است و $A_{i,i}$ زمانی یک می‌شود که بین دو گره i و j یالی موجود باشد، در غیر اینصورت مقدارش صفر است.

پراپ‌فلو[۵۶]: این شاخص مشابه رتبه‌صفحه ریشه‌دار است، اما محلی‌سازی بیشتری روی آن انجام شده‌است. پراپ‌فلو متناسب با احتمال رسیدن یک ولگشت کراندار به y است، که از x شروع می‌شود و از l گام هم بیشتر نیست. این ولگشت پیوندها را بر اساس وزن انتخاب می‌کند و زمانی که به گره y برسد و یا این که مجدداً گره x را ملاقات کند، خاتمه می‌یابد. این روش، عددی را تولید می‌کند که می‌توان به عنوان برآورد احتمال پیوندهای جدید به کار گرفته شود [۳۲]. اگر x و y به طور مستقیم به یکدیگر پیوند داشته‌باشند، پراپ‌فلوی آن‌ها به صورت زیر محاسبه می‌شود:

$$PF(x,y) = PF(a,x)\frac{w_{xy}}{\sum_{k\in\Gamma(x)} w_{xk}} \qquad (۲۱-۲)$$

[۵۴] Rooted PageRank

[۵۵] PageRank

[۵۶] PropFlow

که در آن k همسایهٔ گره x است که عمق آن از نقطه شروع بیشتر از عمق گره x است. w_{xy} بیانگر وزن پیوند بین x و y است و a گرهٔ پیشین x در مسیر ولگشت است. اگر گرهٔ آغازین x باشد، آنگاه: $PF(a,x)=1$. اگر x و y به طور غیرمستقیم با هم مرتبط باشند، $PF(x,y)$ مجموع پراپ‌فلوهای تمام کوتاه‌ترین مسیرها از x به y است. برخلاف رتبه‌صفحه ریشه‌دار، محاسبهٔ پراپ‌فلو به راه‌اندازی مجدد ولگشت و یا همگرایی نیازی ندارد و در عوض به سادگی از یک جستجوی اول سطح محدود شده به ارتفاع l استفاده می‌کند. بنابراین، راهکاری سریع‌تر از سیم‌رنک و رتبه‌صفحه ریشه‌دار است.

۲-۳-۳ روش‌های بیشینه همانندی

یک دسته از روش‌های پیش‌بینی پیوند، دسته روش‌های مبتنی بر تخمین بیشینه همانندی هستند. این روش‌ها ابتدا یک سری اصول ساختاری برای ساختار شبکه در نظر می‌گیرند و با بیشینه کردن شباهت ساختار دیده شده از شبکه، مجموعه قوانین و پارامترهای مشخصی را به دست می‌آورند. سپس، احتمال وجود هر کدام از پیوندهای دیده‌نشده می‌تواند با توجه به این قوانین و پارامترها محاسبه شود.

از نقطه نظر کاربردهای تجربی، یک مشکل اساسی برای روش‌های بیشینه همانندی این است که بسیار زمان‌بر هستند. یک الگوریتم خوب طراحی شده از این نوع، قادر است با شبکه‌هایی با حداکثر چند هزار گره در زمان قابل قبول کار کند، اما وقتی با گراف‌های بسیار بزرگ شبکه‌های اجتماعی برخط که معمولاً از میلیون‌ها گره تشکیل شده‌اند مواجه می‌شود، قادر به انجام عملیات نخواهد بود. علاوه بر این، روش‌های بیشینه همانندی معمولاً جزو دقیق‌ترین و بهترین روش‌های پیش‌بینی پیوند نیستند. با این حال، این روش‌ها دید ارزشمندی از ساختار شبکه برای ما به ارمغان می‌آورند که در روش‌های دیگر مثل روش‌های بر پایهٔ شباهت یا روش‌های احتمالاتی نخواهیم داشت.

۲- ۳- ۴ مدل‌های احتمالاتی

هدف روش‌های احتمالاتی این است که ساختار زیرین یک شبکه را استخراج کند و سپس با استفاده از مدل استخراج‌شده، پیش‌بینی پیوند را انجام دهند. با فرض داشتن یک شبکهٔ هدف مثل $G = (V, E)$، این روش‌ها یک تابع هدف ساخته شده را بهینه می‌کنند تا یک مدل از مجموعهٔ پارامترهای Θ بسازند که بتواند به بهترین شکل بر شبکهٔ هدف مطابق شود. سپس احتمال وجود و یا تشکیل یک پیوند ناموجود مثل (i, j) با احتمال شرطی $P(A_{ij} = 1|\Theta)$ تخمین زده می‌شود. این روش‌ها به سه دستهٔ اصلی مدل احتمالاتی رابطه‌ای[۵۷] یا PRM، مدل احتمالاتی موجودیت-رابطه‌ای[۵۸] یا $PERM$، و مدل تصادفی رابطه‌ای[۵۹] یا SRM تقسیم‌بندی می‌شوند [۳۳].

۲- ۳- ۵ معیارهای مبتنی بر نظریه اجتماعی

شاخص‌هایی که پیش از این بیان شد، صرفا از گره و همبندی استفاده می‌کردند. شاخص‌های پیش‌بینی پیوند که بر پایهٔ نظریهٔ اجتماعی استوار هستند، می‌توانند عملکرد را با گرفتن اطلاعات تعاملات اجتماعی به ویژه در شبکه‌های با مقیاس بزرگ بهبود بخشند[۳۴]. این اطلاعات می‌توانند مفاهیمی همچون روابط قوی[۶۰]، روابط ضعیف[۶۱]، انجمن‌ها و تعادل ساختاری باشد. والورد-ریبازا[۶۲] و لوپز[۶۳]، همبندی و اطلاعات انجمن را با در نظر گرفتن علاقه و رفتارهای کاربران ترکیب کردند و در نهایت پیوندهای آینده در توییتر را پیش‌بینی کردند[۳۵]. این نشان می‌دهد که این روش می‌تواند به شکلی کارآمد در بهبود عملکرد پیش‌بینی پیوند در شبکه‌های اجتماعی در مقیاس بزرگ موثر واقع شود. لیو[۶۴] و همکاران یک مدل پیش‌بینی پیوند بر اساس روابط ضعیف و مرکزیت گره‌ها[۶۵] ارائه کردند. مرکزیت، مبین

[۵۷] Probabilistic Relational Model
[۵۸] Probabilistic Entity Relationship Model
[۵۹] Stochastic Relational Model
[۶۰] Strong ties
[۶۱] Weak ties
[۶۲] Valverde-Rebaza
[۶۳] Lopes
[۶۴] Liu
[۶۵] Node Centeralities

اهمیت و تاثیرگذاری بیشتر در شبکه است و همچنین برای بهبود دقت پیش‌بینی از مفهوم رابطهٔ ضعیف استفاده شده‌است. بنابراین، هر یک از همسایگان مشترک گره‌ها متناسب با مرکزیت خود بر روی پیش‌بینی پیوند اثر خواهند داشت. این مدل به صورت زیر تعریف شده است:

$$LCW(x,y) = \sum_{z}(\omega(z).f(z))^{\beta},\ f(z) = \begin{cases} 1 & \text{if} \quad z \in \Gamma(x) \cap \Gamma(y), \\ 0 & \text{if} \quad otherwise. \end{cases} \qquad (۲۲\text{-}۲)$$

که $\omega(z)$ به میزان مرکزیت هر گره در گراف شبکه اشاره می‌کند.

۲-۳-۶ روش‌های مبتنی بر یادگیری

روش‌های طبقه‌بندی مبتنی بر ویژگی[۶۶]: در گراف $G(V,E)$ که x و y گره‌های آن بوده $(x,y \in V)$ و $l^{(x,y)}$ برچسب زوج گره‌های نمونه (x,y) هستند. در پیش‌بینی پیوند، هر جفت غیرمتصل گره‌ها به یک نمونه شامل برچسب کلاس و ویژگی‌های توصیف زوج گره‌ها تطبیق می‌یابد. بنابراین، یک جفت از گره‌ها می تواند به عنوان مثبت برچسب شود اگر یک پیوند اتصال گره‌ها وجود داشته باشد، در غیر این صورت، به عنوان منفی برچسب می‌شود. برچسب x و y به شرح زیر است:

$$l^{(x,y)} = \begin{cases} +1 & \text{if} \quad (x,y) \in E, \\ -1 & \text{if} \quad (x,y) \notin E \end{cases} \qquad (۲۳\text{-}۲)$$

در این حالت پیش‌بینی پیوند شبیه طبقه‌بندی دودویی است و می‌توان از الگوریتم‌های طبقه‌بندی مانند درخت تصمیم[۶۷]، شبکه عصبی[۶۸]، ماشین بردار پشتیبان[۶۹] و غیره استفاده کرد.

[۶۶] Feature-based Classification

[۶۷] Decision Tree

[۶۸] Neaural Network

[۶۹] Support Vector Machine

روش‌های گراف احتمالاتی: در یک شبکه اجتماعی، به پیوند میان هر زوج گره می‌توان مقدار احتمالی را نسبت داد مانند شباهت همبندی یا احتمال انتقال در ولگشت. این روش‌ها یک مدل احتمالاتی ایجاد می‌کنند و با کمک یال‌هایی که قبلا مشاهده شده‌اند، پارامتر لازم را تنظیم کرده و به پیش‌بینی پیوند می‌پردازند. این مدل‌ها از ویژگی‌های گره‌ها و یال‌ها به منظور مدل‌سازی توزیع احتمال توأم موجودیت‌ها و یال‌های میان آن‌ها استفاده می‌کنند.

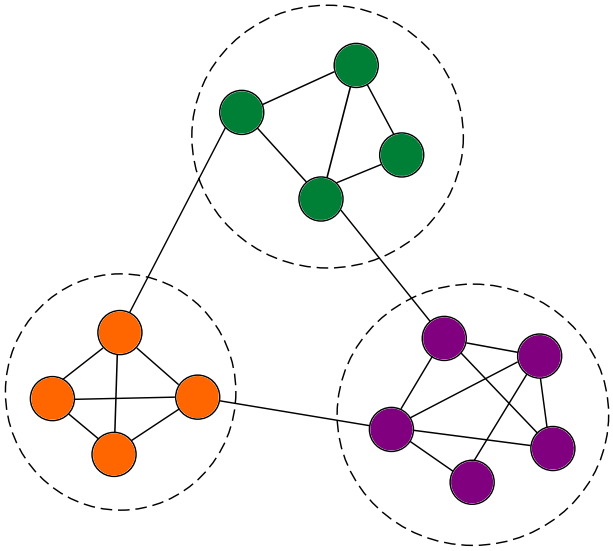
روش‌های تجزیه ماتریس: منون [۷۰] و همکاران پیش‌بینی پیوند را به مانند یک مساله تکمیل ماتریس معرفی کردند و روش تجزیه به عامل‌های ماتریس را به منظور پیش‌بینی پیوند گسترش دادند[۳۶].

۲-۴- مروری بر روش‌های تشخیص انجمن

ساختار انجمن‌های یک شبکه یکی از مهمترین ویژگی‌های ساختاری آن شبکه است. به عنوان یک تعریف کیفی، می‌توان گفت که انجمن‌ها زیرمجموعه‌هایی از گره‌های یک شبکه هستند که ارتباطات در درون آن‌ها به نسبت بیرون، بیشتر و در اصطلاح چگال‌تر است [۳۷]. در تعریف دیگری، انجمن‌ها گروهی از گره‌ها خوانده می‌شوند که احتمالاً ویژگی‌های مشترک دارند و یا نقش‌های مشابهی در شبکه ایفا می‌کنند [۳۸]. برای مثال در یک شبکهٔ اجتماعی، دانشجویان هم‌رشته در یک دانشگاه و در مقاطع نزدیک، ارتباطات بیشتری با هم دارند و در واقع تشکیل یک انجمن می‌دهند. همین امر باعث می‌شود که اعضای یک انجمن ویژگی‌های شبیه به هم داشته باشند و علایق و سلایق آن‌ها به هم نزدیک باشد و رفتارهای مشابهی از خود نشان دهند. همچنین تاثیرگذاری اعضای یک انجمن روی بقیهٔ اعضای آن انجمن بیشتر است. مطالعهٔ انجمن‌ها به دلایل متعددی اهمیت دارد. برای مثال می‌توان از مطالعهٔ انجمن‌های یک شبکه اجتماعی، افرادی با سلیقه مشترک را شناسایی کرد و با استفاده از این کار، اهداف اجتماعی، سیاسی، تجاری و... خود را دنبال کرد.

[۷۰] Menon

در شکل ۲- ۱ می‌توان یک گراف ساده با سه انجمن را مشاهده کرد که با خط‌چین از هم جدا شده‌اند. همان‌طور که در این شکل نیز مشخص است، تعداد چگالی یال‌های درون انجمن‌ها بیشتر از بیرون آن‌هاست و گره‌ها درون انجمن‌ها روابط نزدیک‌تری به هم دارند.



شکل ۲- ۱: یک شبکهٔ نمونه با سه انجمن

در این بخش، ابتدا مروری خواهد شد بر روش‌های تشخیص انجمن و دسته‌بندی از آن‌ها ارائه خواهد شد. سپس روش تشخیص انجمن مد نظر این پژوهش یعنی الگوریتم نقشه‌اطلاعات[۷۱] با جزییات بیشتری بررسی خواهد شد.

۲- ۴- ۱ انواع روش‌های تشخیص انجمن

روش‌های تشخیص انجمن به دسته‌های مختلفی طبقه‌بندی می‌شوند و طبقه‌بندی‌های مختلفی توسط پژوهشگران مختلف از این الگوریتم‌ها ارائه شده‌است [۳۷] [۳۸] [۳۹]. در این بخش به مرور تعدادی از این روش‌ها می‌پردازیم.

[۷۱] Infomap

روش‌های سنتی

این روش‌ها خود به چند دسته تقسیم‌بندی می‌شوند:

روش‌های افراز گراف[۷۲]: این روش‌ها سعی می‌کنند یک گراف را به تعداد مشخصی زیرگروه با سایز مشخص تقسیم کنند، به طوری که تعداد یال‌هایی که بین زیرگروه‌ها قرار می‌گیرند کمینه باشد. به تعداد یال‌هایی که بین زیرگروه‌ها قرار می‌گیرند اندازهٔ برش[۷۳] گفته می‌شود. در این روش‌ها تعداد زیرگروه‌ها اهمیت دارد، چون اگه این پارامتر آزاد باشد، در نهایت به یک پاسخ بدیهی خواهیم رسید که همهٔ گره‌ها در یک زیرگروه قرار بگیرند. همچنین اندازهٔ زیرگروه‌ها نیز مهم است چون در صورت آزاد بودن این پارامتر نیز به یک پاسخ بدیهی دیگر خواهیم رسید که گرهی با کمترین درجه به عنوان یک زیرگروه و سایر گره‌ها به عنوان زیرگروه دیگر معرفی شوند.

روش‌های خوشه‌بندی سلسله‌مراتبی[۷۴]: چون در حالت کلی اطلاعات بسیار کمی دربارهٔ تعداد خوشه‌های یک گراف وجود دارد، تعیین تعداد و اندازهٔ زیرگروه‌ها کار ساده‌ای نیست. از طرف دیگر گراف‌ها معمولن ساختاری سلسله‌مراتبی[۷۵] دارند به این معنی که سطح‌های مختلفی از گروه‌بندی را شامل می‌شوند و انجمن‌های کوچکتر در داخل انجمن‌های بزرگ‌تر قرار می‌گیرند. نقطه شروع این روش‌ها تعیین یک معیار شباهت بین گره‌هاست. وقتی این معیار شباهت انتخاب شد، برای هر جفت گره محاسبه می‌شود (صرف نظر از این که گره‌ها به هم متصل هستند یا نه) و در نهایت، هدف، رسیدن به خوشه‌بندی‌ایست که شباهت داخل گروه‌ها را بیشینه کند. خود این روش‌هابه دو دسته تقسیم می‌شوند:

۱. **الگوریتم‌های تجمیعی**[۷۶] که رویکردی پایین به بالا[۷۷] دارند. در ابتدا هر گره به تنهایی یک خوشه را تشکیل می‌دهد و سپس خوشه‌هایی با بیشترین شباهت با هم ترکیب شده و خوشه‌های بزرگتر را می‌سازند

[۷۲] Graph Partitioning
[۷۳] cut size
[۷۴] Heirarchical Clustering
[۷۵] Heirarchical
[۷۶] Agglomerative
[۷۷] bottom-up

و این کار تا وقتی ادامه پیدا می‌کند که یک خوشه که همهٔ گره‌ها را در بر دارد بماند.

۲. **الگوریتم‌های تقسیمی**[۷۸] که رویکردی بالا به پایین[۷۹] دارند. در ابتدا کل گره‌های گراف یک خوشه را تشکیل می‌دهند و در ادامه این خوشه به صورت تکراری از روی یال‌هایی که کمترین شباهت را دارند شکسته می‌شوند تا در نهایت هر گره در یک خوشه قرار گیرد.

از مشکلات این روش‌ها می‌توان به هزینهٔ محاسباتی بالا، وابستگی زیاد به معیار شباهت انتخاب شده، کارایی نه چندان خوب در گراف‌هایی بدون ساختار سلسله‌مراتبی، و... اشاره کرد.

روش‌های خوشه‌بندی افرازی[۸۰]: این روش‌ها نیز همان‌طور که از اسمشان پیداست، سعی می‌کنند که مجموعه داده‌ها را که همان گره‌های گراف است خوشه‌بندی کنند. در این‌جا نیز تعداد خوشه‌ها از قبل مشخص است. همچنین می‌بایست گره‌های گراف را به فضایی برد که در آن قابلیت اندازه‌گیری وجود داشته باشد و بتوان فاصلهٔ بین گره‌ها را محاسبه کرد. سپس روش‌هایی مانند k-means و سایر روش‌های این خانواده را روی آن‌ها اعمال کرد.

روش‌های خوشه‌بندی طیفی[۸۱]: در این دسته از روش‌ها نیاز است که یک ماتریس فاصله (مثل S) متناظر با گراف تولید شود که برای هر جفت از گره‌ها، فاصلهٔ آن‌ها را مشخص می‌کند. سپس از روش‌هایی کمک گرفته می‌شود که این ماتریس را با استفاده از بردارویژهٔ ماتریس S (یا دیگر ماتریس‌هایی که با استفاده از آن به دست می‌آیند)، به خوشه‌هایی افزار کند.

[۷۸] Divisive

[۷۹] top-down

[۸۰] Partitional clustering

[۸۱] Spectral Clustering

روش‌های بر پایهٔ پیمانگی

یک روش تشخیص انجمن خوب روشی‌است که افراز خوبی انجام دهد. اما برای تعریف یک افراز خوب نیاز به یک معیار کمی داریم. یکی از مشهورترین این معیارها، پیمانگی[۸۲] است. این معیار توسط نیومن[۸۳] و گیروَن[۸۴] در سال ۲۰۰۴ معرفی شد[۴۰]. این معیار بر پایهٔ این ایده بنا شده‌است که از یک گراف تصادفی انتظار نمی‌رود که ساختار خوشه‌ای مشخصی داشته باشد. در نتیجه می‌توان با مقایسهٔ چگالی یال‌های زیرگراف‌های به دست آمده با زیرگراف‌های یک گراف تصادفی با توزیع درجات یکسان، وجود ساختار خوشه‌ها در گراف را مشخص کرد. رابطهٔ ارائه شده برای محاسبهٔ این معیار به صورت زیر است:

$$Q = \frac{1}{2m}\sum_{ij}(A_{ij} - P_{ij})\delta(C_i, C_j) \qquad (۲۴-۲)$$

که در آن مجموع، روی تمام جفت گره‌ها اعمال می‌شود. در رابطهٔ ۲- ۲۴، A ماتریس مجاورت، m تعداد کل یال‌ها و P_{ij} احتمال وجود یال بین دو گرهٔ i و j است. تابع $\delta(i,j)$ نیز مشخص می‌کند که آیا دو گره i و j در یک انجمن هستند یا خیر، به این صورت که اگر عضو یک انجمن بودند مقدار 1 و در غیر این صورت مقدار 0 به خود می‌گیرد. در نهایت مقدار Q عددی بین $+1$ و -1 خواهد شد که هر چه بزرگتر بودن آن نشان‌دهندهٔ بهتر بودن خوشه‌بندی خواهد بود.

در این دسته از روش‌ها، هدف بیشینه کردن مقدار Q است. در نتیجه این روش‌ها، مسئلهٔ تشخیص انجمن را تبدیل به یک مسئلهٔ بهینه سازی می‌کنند. حل این مسائل بهینه‌سازی از روش‌های مختلفی مثل روش حریصانه[۸۵]، روش تبرید شبیه‌سازی‌شده[۸۶] و غیره قابل حل است. از مشکلات این دسته روش‌ها می‌توان به افزایش سریع فضای مسئلهٔ بهینه‌سازی اشاره کرد که این روش‌ها را برای گراف‌های بزرگ نامناسب می‌کند.

[۸۲] Modularity
[۸۳] Newman
[۸۴] Girvan
[۸۵] greedy
[۸۶] simulated annealing

سایر روش‌ها

دسته‌هایی دیگر از روش‌ها نیز مانند روش‌های استنتاج آماری[۸۷]، الگوریتم‌های پویا[۸۸] و... وجود دارند که توضیح آنها از حوصلهٔ این پژوهش خارج است. اما روشی که در این پژوهش مد نظر است، الگوریتم نقشه‌اطلاعات است که در دستهٔ روش‌های پویا دسته‌بندی می‌شود. در بخش بعد به توضیح در مورد این الگوریتم می‌پردازیم.

۲-۴-۲ الگوریتم نقشه‌اطلاعات

این الگوریتم در سال ۲۰۰۸ توسط روسوَل[۸۹] و برگستروم[۹۰] [۴۱] به عنوان روشی برای تشخیص انجمن‌ها معرفی شد. همان‌طور که اشاره شد، این روش یک روش پویاست. در این روش سعی می‌شود که اطلاعات یک فرآیند پویا که روی یک گراف در حال شکل‌گیری است، طوری کد شود که به بهترین شکل فشرده شود. این فرآیند پویا یک ولگشت با طول بی‌نهایت است. در واقع این مسئله با بهینه‌سازی یک تابع هزینه به نام حداقل طول توصیف[۹۱] به دست می‌آید[۴۲]. این بهینه‌سازی با ترکیبی از جستجوی حریصانه و تبرید شبیه‌سازی‌شده قابل حل است.

برای روشن‌تر شدن موضوع فرض کنید می‌خواهیم ولگشت تصویر شده در شکل ۲- ۲آ را کد کنیم. کدگذاری هافمن[۹۲] می‌تواند این کار را برای ما انجام دهد. با استفاده از این روش، همان‌طور که در شکل ۲- ۲ب به هر کدام از گره‌ها یک دنباله یکتا از ۰ و ۱ نسبت می‌دهیم، سپس برای کد کردن مسیر، شمارهٔ گره‌های دیده‌شده در مسیر را به ترتیب پشت سر هم قرار می‌دهیم. طول کد به دست آمده برای این مسیر ۳۱۴ بیت خواهد بود. اما یک روش دیگر برای کدگذاری این است که از یک کدگذاری دوسطحی استفاده کنیم، به این صورت که به هر کدام از خوشه‌های اصلی گراف یک کد یکتا اختصاص بدهیم، اما کد گره‌ها داخل خوشه‌ها می‌تواند مجدداً استفاده شود. این کدها در

[۸۷] statistical inference

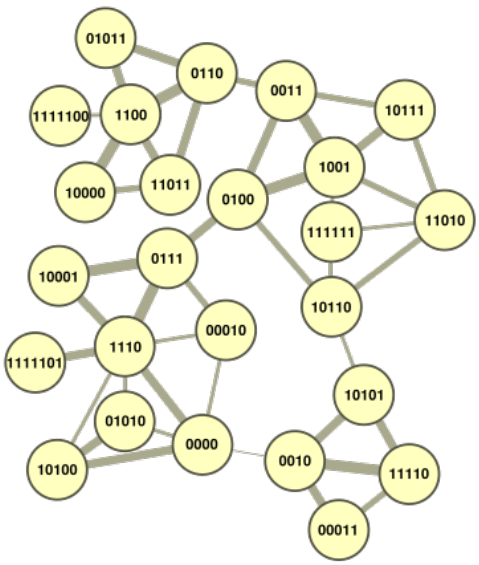
[۸۸] dynamic algorithms

[۸۹] Rosvall

[۹۰] Bergestrom

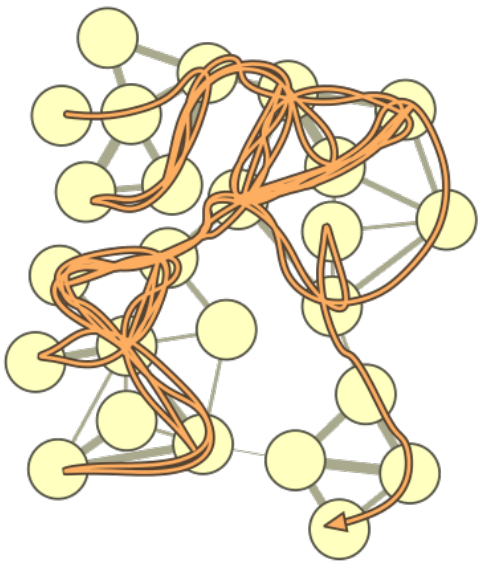
[۹۱] Minimim Description Length

[۹۲] Huffman Coding

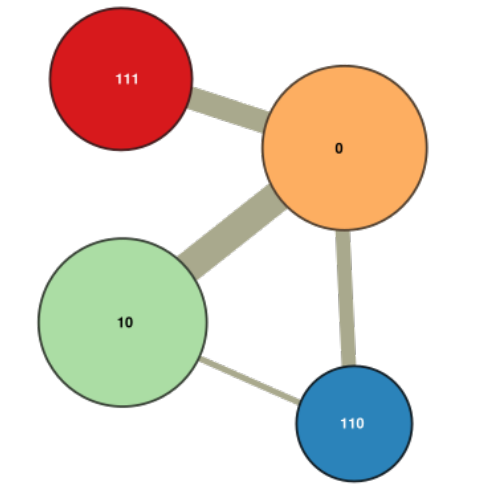


1111100 1100 0110 11011 10000 11011 0110 0011 10111 1001 0011
1001 0100 0111 10001 1110 0111 10001 0111 1110 0000 1110 10001
0111 1110 0111 1110 1111101 1110 0000 10100 0000 1110 10001 0111
0100 10110 11010 10111 1001 0100 1001 10111 1001 0100 1001 0100
0011 0100 0011 0110 11011 0110 0011 0100 1001 10111 0011 0100
0111 10001 1110 10001 0111 0100 10110 111111 10110 10101 11110
00011

(ب)

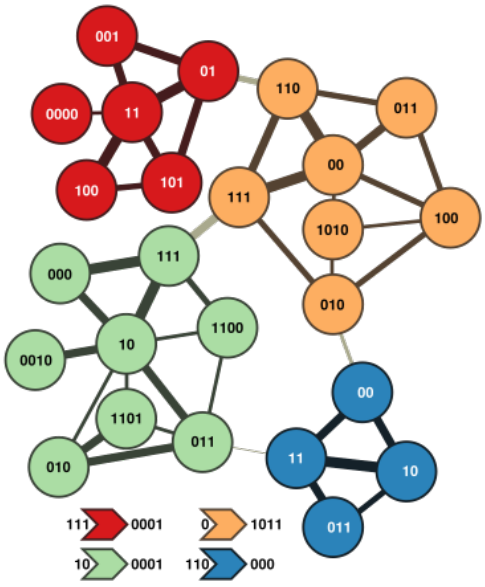


(آ)



111 0000 11 01 101 100 101 01 0001 0 110 011 00 110 00 111 1011 10
111 000 10 111 000 111 10 011 10 000 111 10 111 10 0010 10 011 010
011 10 000 111 0001 0 111 010 100 011 00 111 00 011 00 111 00 111
110 111 110 1011 111 01 101 01 0001 0 110 111 00 011 110 111 1011
10 111 000 10 000 111 0001 0 111 010 1010 010 1011 110 00 10 011

(د)



(ج)

شکل ۲- ۲: چگونگی عملکرد الگوریتم نقشه‌اطلاعات

(تصاویر برگرفته از [۴۳])

شکل‌های ۲- ۲ج و ۲- ۲د قابل مشاهده هستند. همچنین کدهای ورود و خروج از هر خوشه نیز در شکل ۲- ۲ج به ترتیب قبل و بعد از فلش‌ها نوشته شده‌اند. با استفاده از این کدهای جدید، می‌توانیم مسیر مورد نظر را با استفاده از ۲۴۳ بیت کد کنیم. این مسیر در پایین شکل ۲- ۲ج نمایش داده شده‌است. همان‌طور که مشاهده می‌شود، با استفاده از این کدگذاری جدید توانستیم ٪۳۲ از طول کد نهایی بکاهیم.

اما این کدگذاری را می‌بایست به دست آوریم. برای این کار فرض کنید یک افراز مثل M داریم که n گره $\alpha = 1, \ldots, n$ را به m زیرمجموعه $i = 1, \ldots, m$ افراز می‌کند. حد پایین طول کد به دست آمده توسط M را به شکل تابع $L(\mathsf{M})$ با توجه به نظریهٔ شانون تعریف می‌کنیم. تابع مورد نظر به شکل زیر تعریف می‌شود: جزییات کامل‌تر دربارهٔ این الگوریتم در [۴۳] مورد بحث قرار گرفته‌است. همچین دربارهٔ نسخهٔ وزن‌دار این الگوریتم نیز صحبت شده است که از آن در فصل ۳ که روش پیشنهادی این پژوهش تشریح خواهد شد، استفاده شده است.

فصل ۳: روش پیشنهادی

۳- ۱- مقدمه

در این فصل به ارائهٔ روش پیشنهادی برای پیش‌بینی پیوند پرداخته می‌شود. در ابتدا ایدهٔ اولیه را مطرح می‌کنیم که همان بهره گرفتن از اطلاعات انجمن‌ها در کمک به بهبود کارایی روش‌های پیش‌بینی پیوند است. اشاره‌ای به کارهای گذشته در این زمینه خواهد شد. سپس ایدهٔ اصلی این پژوهش به طور کامل شرح داده خواهد شد. در بخش بعد به تعریف ریاضی روش پیشنهادی پرداخته خواهد شد و فرمول‌بندی آن را توضیح داده می‌شود. سپس استفاده از وزن یال‌ها و نحوهٔ ترکیب آن با روش پیشنهادی مورد بررسی قرار خواهد گرفت. در آخر نیز به یک رویکرد عملی برای محاسبهٔ روش پیشنهادی پرداخته خواهد شد و دربارهٔ ویژگی‌های این رویکرد بحث خواهد شد.

۳- ۲- پیش‌بینی پیوند داخل انجمن‌ها

ایدهٔ استفاده از اطلاعات انجمن‌ها برای کمک به روش‌های پیش‌بینی پیوند، پیش از این در کار ساندراجان[۱] و هاپکرفت[۲] [۴۴] استفاده شده بود. آن‌ها تعریف شاخص‌های شباهت را تغییر داده بودند و اطلاعات انجمن‌ها را در آن وارد کرده بودند و نتیجه گرفته بودند که در بیشتر مواقع، این شاخص‌های جدید گسترش‌یافته، می‌توانند از شاخص‌های معمولی کارایی بهتری از خود نشان دهند.

در این پژوهش، ایده اصلی این است که در بیشتر موارد، پیوندهای بالقوه درون انجمن‌ها بسیار کمتر از پیوندهای بالقوه بین انجمن‌ها هستند. این موضوع با یک مثال ساده روشن‌تر می‌شود. فرض کنید یک شبکه داریم که از ۱۰۰ گره تشکیل شده‌است. این شبکه درون خود دارای ۵ انجمن است که هر کدام از این انجمن‌ها ۲۰ گره درون خود دارند. بنابر این تعداد کل پیوندهای بالقوه درون کل انجمن‌ها این شبکه برابر خواهد بود با ۱۹×۲۰×۵ که برابر است با ۱۹۰۰ یال بالقوه. اما از سوی دیگر تعداد کل پیوندهای بالقوه بین انجمن‌ها برابر خواهد بود با ۸۰×۲۰×۵ که برابر است با ۸۰۰۰ یال بالقوه. همان‌طور که مشاهده می‌شود، تعداد یال‌های بالقوه داخل انجمن‌ها بسیار کمتر از تعداد یال‌های بالقوه بین انجمن‌ها خواهد بود. همچنین، در بیشتر موارد، چگالی یال‌های داخل یک انجمن بیشتر از چگالی یال‌های بیرون از انجمن‌ها است، به این دلیلِ بدیهی که تعریف «انجمن» این‌گونه حکم می‌کند. یعی برای مثال درصد بسیار بیشتری از ۱۹۰۰ یال بالقوه یادشده داخل انجمن‌ها در واقع موجودند تا ۸۰۰۰ یال بالقوه بیرون از انجمن‌ها.

این ایده‌ها کمک می‌کنند تا بتوان به روش مناسبی دست پیدا کرد. بنا بر بحث‌های بالا، محدود کردن پیش‌بینی پیوند به پیش‌بینی درون انجمن‌ها، کمک خواهد کرد تا بتوان پیوندهای درست بیشتری را تشخیص داد. دلیل این امر این است که یک گره، با احتمال بیشتری با یک گره از انجمن خود پیوند برقرار می‌کند تا یک گره از بیرون انجمن

[۱] Soundarajan

[۲] Hopcroft

خود. البته طبیعی است که با این کار پیش‌بینی پیوندهای بالقوه بین انجمن‌ها را از دست خواهیم داد، اما در تعداد محدودی پیش‌بینی، این کار می‌تواند به افزایش دقت پیش‌بینی‌های ما کمک شایانی بکند.

۳- ۲- ۱ تعریف ریاضی روش

برای ارائه تعریف ریاضی دقیق‌تر برای این روش‌ها، فرض کنید یک شبکه داریم که قصد داریم در آن پیش‌بینی پیوند انجام دهیم. ابتدا شاخص شباهت مورد نظر خود (برای مثال شاخص همسایه‌های مشترک) را روی شبکه محاسبه می‌کنیم. همچنین به طور موازی یک روش تشخیص انجمن را نیز بر روی شبکه مورد نظر خود اجرا می‌کنیم تا انجمن‌های موجود در شبکه را بشناسیم. فرض کنید که $SM(x,y)$ مقدار شاخص شباهت مورد نظر بین دو گره x و y باشد. رابطهٔ محاسبهٔ روش پیشنهادی به صورت زیر خواهد بود:

$$SM'(x,y) = SM(x,y) \times CO(x,y), \qquad (۱-۳)$$

که در آن، CO به صورت زیر تعریف می‌شود:

$$CO(x,y) = \begin{cases} 1 & \text{if} \quad comm(x) = comm(y), \\ 0 & \text{if} \quad otherwise. \end{cases} \qquad (۲-۳)$$

و مقدار $comm(x)$ به انجمنی اشاره دارد که شامل گره x می‌شود.

همان‌طور که از رابطه ۳- ۱ مشخص است، پس از محاسبهٔ شاخص شباهت و تشخیص انجمن‌های شبکهٔ مورد نظر، مقدار نهایی روش پیشنهادی یعنی SM'، شاخص شباهت اصلی را فقط بین گره‌هایی که در یک انجمن یکسان حضور دارند مورد توجه قرار می‌دهد و بین بقیهٔ جفت گره‌هایی که در یک انجمن یکسان نیستند، مقدار صفر به خود می‌گیرند. بنابراین در این‌جا در واقع پیش‌بینی پیوند داخل انجمن‌ها انجام می‌گیرد که در بخش قبل نیز به آن اشاره

شد.

۳- ۲- ۲ استفاده از وزن پیوندها

همان‌طور که در بخش‌های قبل نیز به آن اشاره شد، در بسیاری موارد اطلاعات مفید دیگری نیز در شبکه‌ها وجود دارند که می‌توانند به بهبود کارایی و افزایش دقت روش‌های پیش‌بینی پیوند کمک کنند. یکی از این اطلاعات ارزشمند، وزن پیوندهاست. همان‌طور که از عنوان این پژوهش نیز بر می‌آید، قصد آن این است که بتواند روش‌های پیش‌بینی پیوند را با کمک گرفتن از وزن پیوندها بهبود ببخشد و دقت پیش‌بینی‌ها را افزایش دهد.

با توجه روشی که در بخش گذشته پیشنهاد شد، این روش از دو گام مجزا تشکیل شده‌است. گام اول محاسبهٔ معیارهای شباهت بین هر دو گره دلخواه در شبکه که پیوندی بین آن دو وجود ندارد؛ و گام دوم تشخیص انجمن‌ها در همان شبکه، که در نهایت با ترکیب این دو گام نتیجهٔ نهایی به دست می‌آید. بر طبق توضیحاتی که در قسمت‌های قبل داده شد، هر دوی این گام‌ها می‌توانند از وزن پیوندها استفاده کنند. در گام اول یعنی محاسبهٔ معیارهای شباهت، در بخش ۲- ۳- ۲ برای هر کدام از شاخص‌ها گسترش آن‌ها به حالت وزن‌دار نیز بررسی شد و با استفاده از روابط موجود می‌توان آن‌ها را در حالت وزن‌دار محاسبه کرد. در گام دوم یعنی تشخیص انجمن‌ها نیز همان‌طور که گفته شد، الگوریتم نقشه‌اطلاعات دارای حالت وزن‌دار است و می‌تواند از شبکه ورودی وزن‌دار استفاده کرده و انجمن‌ها را با استفاده از آن تشخیص دهد.

همان‌طور که عنوان شد، در روش پیشنهادی دو گام محاسبهٔ شاخص و تشخیص انجمن وجود دارد که هر بخش می‌تواند بدون وزن یا وزن‌دار انجام گیرد. در نتیجه با توجه به این که در هر کدام از این دو بخش، کدام حالت (یعنی بدون وزن یا وزن‌دار) را انتخاب کنیم، روش‌های پیشنهادی به چهار روش گسترش داده می‌شود:

۱. محاسبهٔ شاخص بدون وزن همراه با تشخیص انجمن بدون وزن
۲. محاسبهٔ شاخص بدون وزن همراه با تشخیص انجمن وزن‌دار
۳. محاسبهٔ شاخص وزن‌دار همراه با تشخیص انجمن بدون وزن

۴. محاسبهٔ شاخص وزن‌دار همراه با تشخیص انجمن وزن‌دار

جدول ۳- ۱ این تقسیم‌بندی را نشان می‌دهد. همان‌طور که در این جدول نشان داده شده‌است، از این پس برای ارجاع دادن به هر کدام از این روش‌ها از علائم اختصاری آن‌ها استفاده می‌کنیم که به ترتیب عبارتند از: UU، UW، WU و WW.

جدول ۳- ۱: چهار روش پیشنهادی برای استفاده از اطلاعات وزن پیوندها، و انجمن‌ها

		شاخص پیش‌بینی پیوند	
		بدون وزن	وزن‌دار
روش تشخیص انجمن	بدون وزن	ترکیب شاخص بدون وزن با تشخیص انجمن بدون وزن (شماره ۱) که با UU نمایش داده می‌شود	ترکیب شاخص وزن‌دار با تشخیص انجمن بدون وزن (شماره ۳) که با WU نمایش داده می‌شود
	وزن‌دار	ترکیب شاخص بدون وزن با تشخیص انجمن وزن‌دار (شماره ۲) که با UW نمایش داده می‌شود	ترکیب شاخص وزن‌دار با تشخیص انجمن وزن‌دار (شماره ۴) که با WU نمایش داده می‌شود

این چهار روش هر کدام می‌توانند با توجه به ساختار و ویژگی‌های شبکه‌های مختلف، در مجموعه داده‌های مختلف، متفاوت عمل کنند و کارایی‌های مختلفی از خود نشان دهند. هدف این پژوهش این است که این چهار روش را در حالت‌های مختلف شبکه‌ها بررسی کند و نشان دهد که هر کدام از آن‌ها در چه نوع شبکه‌هایی عملکرد خوبی از خود نشان دهند و از سوی دیگر در کدام نوع شبکه‌ها عملکرد مناسبی ندارند. همچنین دلیل عملکردهای متفاوت این روش‌ها نیز بررسی خواهد شد. برای مثال اگر در شبکه‌ای پیوندهای بیرون انجمن‌ها قوی‌تر باشد، ممکن است روش تشخیص انجمن وزن‌دار به نتایج مناسبی ارائه ندهد و دقت پیش‌بینی را کم کند. به این موضوع در فصل بعد به تفصیل پرداخته خواهد شد.

۳- ۲- ۳ رویکرد عملی محاسبهٔ شاخص‌های پیشنهادی

با وجود این که در بخش قبل یک تعریف ریاضی از نحوهٔ محاسبهٔ روش‌ها ارائه شد، اما می‌توان از رویکرد دیگری نیز برای محاسبهٔ آن‌ها استفاده کرد. البته باید توجه داشت که نتایج این رویکرد جدید محاسبه ممکن است در مواردی با نتایج رویکرد ریاضی که در رابطهٔ ۳- ۱ معرفی شد متفاوت باشد. در رویکرد قبلی، گام تشخیص انجمن به طور موازی با گام محاسبهٔ شاخص شباهت انجام می‌شد و در نهایت نتایج این دو گام با هم ترکیب می‌شدند. اما در این رویکرد گام تشخیص انجمن می‌تواند ابتدا انجام شود، سپس شبکه مورد نظر به انجمن‌های خود شکسته شود و در گام بعد، برای هر انجمن محاسبهٔ شاخص شباهت به طور مجزا انجام گیرد. برتری این رویکرد نسبت به رویکرد قبلی در این نکته است که مرتبهٔ[۳] انجمن‌های یک شبکه به مراتب از خود شبکه کمتر است. با توجه به این که پیچیدگی زمانی[۴] برای بیشتر روش‌ها طبق آن‌چه که در [۱۱] عنوان شده است، از $O(n^2)$ (و ضرایب آن) است[۵]، این امر باعث می‌شود که از پیچیدگی زمانی فرآیند محاسبهٔ شاخص‌های شباهت کاسته شود.

برای روشن‌تر شدن موضوع یک مثال بسیار ساده ارائه می‌شود: فرض کنید یک گراف با ۵۰۰۰ گره داریم که دارای ۱۰۰ انجمن با میانگین تعداد اعضای ۵۰ است. برای محاسبهٔ هر یک از شاخص‌های شباهت (که پیچیدگی زمانی از مرتبه $O(n^2)$ دارند) برای مثال معیار «همسایه‌های مشترک» لازم است که محاسباتی از مرتبهٔ $5\,000^2 = 25\,000\,000$ انجام شود. اما اگر از رویکرد دوم برای محاسبه استفاده کنیم، محاسبات ما از مرتبهٔ $100 \times 50^2 = 250\,000$ خواهد بود که بسیار کمتر از روش محاسبهٔ اول است.

البته همان‌طور که گفته شد باید توجه داشت که خروجی در این حالت با خروجی حالت قبل ممکن است متفاوت باشد. دلیل این امر نیز این است که اگر دو گره که عضو یک انجمن هستند همسایهٔ مشترکی خارج از انجمن داشته باشند، در معیارهایی که بر پایهٔ همسایه‌های مشترک استوارند (مانند همسایه‌های مشترک، آدامیک/ادار،

---

[۳] order

[۴] time complexity

[۵] به جز شاخص «وابستگی ترجیحی» که پیچیدگی زمانی آن از مرتبهٔ $O(n)$ است.

تخصیص منابع و...)، با استفاده از رویکرد جدید، این همسایهٔ مشترک محاسبه نمی‌شود و همین امر منشأ تفاوت بین خروجی‌ها خواهد بود. اما در روشی مثل وابستگی ترجیحی این مشکل وجود نخواهد داشت و نتیجه در دو حالت یکسان خواهد بود.

در نهایت به عنوان یک جمع‌بندی دربارهٔ این رویکرد محاسبه، می‌توان گفت که برتری آن نسبت به رویکرد اولیه، کاهش دادن پیچیدگی زمانی و هزینهٔ محاسبات است که موضوع مهمی در پیش‌بینی پیوند محسوب می‌شود. از طرف دیگر کاستی این رویکرد این است که دقیق نیست و ممکن است در برخی موارد با روش اصلی اختلاف داشته باشد.

۳-۳- جمع‌بندی

در این فصل روش‌های پیشنهادی این پژوهش ارائه شدند. ایدهٔ اصلی این روش‌ها، پیش‌بینی پیوند درون انجمن‌ها بود و دلیل این کار نیز چگالی بالای یال‌های درون انجمن‌ها، در کنار پایین‌تر بودن تعداد یال‌های بالقوه درون انجمن‌ها عنوان شد. پس از آن فرمول‌بندی ریاضی برای این روش‌ها عنوان شد. سپس نحوهٔ استفاده از وزن یال‌ها در روش پیشنهادی عنوان شد که در نهایت روش کلی را به چهار روش توسعه دادند. در پایان نیز یک رویکرد عملی اما تقریبی ارائه شد و مزایا و معایب آن بررسی شد. در ادامهٔ پژوهش، این روش‌ها روش مجموعه داده‌هایی که معرفی خواهند شد، اعمال می‌شوند و نتایج به دست آمده از آنها مورد تحلیل و بررسی قرار خواهد گرفت.

فصل ۴: آزمایش‌ها و نتایج و تفسیر آن‌ها

۴-۱- مقدمه

در این بخش ابتدا مجموعه داده‌های مورد استفاده در این پژوهش معرفی می‌شوند و خصوصیات آن‌ها بیان می‌گردد. پس از آن معیارهای ارزیابی نتایج که به طور معمول در مسئلهٔ پیش‌بینی پیوند استفاده می‌شوند مورد بررسی قرار خواهند گرفت. پس از روش‌های پیشنهادی روی مجموعه داده‌ها اعمال شده و نتایج انجام آزمایشات گزارش خواهند شد و این نتایج به دست آمده تحلیل و بررسی خواهند شد. در نهایت نیز یک جمع‌بندی از مطالب این فصل ارائه خواهد شد.

۴-۲- معرفی مجموعه داده‌ها

هدف از این پژوهش، بررسی کارایی روش‌های پیشنهادی با توجه به پارامترهای مختلف در یک نوع شبکهٔ مصنوعی[۱] است. مجموعه داده‌های آزمایشی در این پژوهش، یک نوع شبکهٔ مصنوعی موسوم به شبکه‌های LFR هستند. این شبکه‌ها در اصل به عنوان نوعی شبکهٔ محک[۲] برای روش‌های تشخیص انجمن‌ها توسط لانچیکینِتی[۳]، فُرتوناتو[۴] و رادیکی[۵] در سال ۲۰۰۸ معرفی شدند [۴۵]. شبکه‌های LFR از خانوادهٔ شبکه‌های مقیاس آزاد هستند و در آن‌ها فرض می‌شود که توزیع‌های درجهٔ گره‌ها، قدرت گره‌ها و اندازه انجمن‌ها از توزیع توانی با پارامترهای γ و β تبعیت می‌کنند. این شبکه‌ها، شبکه‌هایی مصنوعی هستند که با تعدادی پارامتر ورودی ساخته می‌شوند و انجمن‌های موجود در شبکه را نیز می‌دهد و می‌توان از آن برای مقایسه روش‌های تشخیص انجمن‌ها استفاده کرد. اما در این پژوهش استفاده‌ای که از این شبکه‌ها انجام خواهد گرفت، در مسئلهٔ پیش‌بینی پیوند است. به این صورت که تاثیر پارامترهای ورودی این شبکه‌ها روی دقت پیش‌بینی مورد بررسی قرار خواهد گرفت. ابتدا توضیحاتی در مورد این شبکه‌ها و پارامترهای ورودی آن‌ها داده خواهد شد. همان‌طور که اشاره شد، این شبکه‌ها می‌توانند با هم مجموعه دلخواه از پارامترها ساخته شده و مورد استفاده قرار گیرند. پارامترهایی که می‌توان برای این شبکه‌ها در نظر گرفت پارامترهایی نظیر تعداد گره‌ها، میانگین درجه گره‌ها، بیشترین اندازهٔ هر انجمن و... هستند. در ادامه لیستی از این پارامترها آورده و توضیح داده خواهند شد:

- N: اولین پارامتر برای تولید این شبکه‌ها، تعداد گره‌ها یا N است.
- k: پارامتر مهم بعدی k یا میانگین درجات گره‌های گراف است.
- $maxk$: پارامتر $maxk$ حداکثر درجهٔ گره‌های گراف را مشخص می‌کند.

[۱] synthesis network
[۲] benchmark
[۳] Lancichinetti
[۴] Fortunato
[۵] Radicchi

- μ_t: پارامتر بسیار مهم بعدی یعنی μ_t که به آن پارامتر تنظیم‌کنندهٔ هم‌بندی[۶] می‌گویند، نسبت یال‌های درون انجمن به یال‌های بیرون انجمن را کنترل می‌کند به این صورت که نسبت یال‌هایی که بین انجمن‌ها قرار دارند به کل یال‌ها، برابر با μ_t و در نتیجه نسبت یال‌های که درون انجمن‌ها قرار دارند برابر با $1 - \mu_t$ خواهد بود.

- μ_w: همانند μ_t پارامتر دیگری نیز وجود دارد که توزیع وزن‌ها را کنترل می‌کند. این پارامتر که به آن پارامتر تنظیم‌کنندهٔ وزن[۷] می‌گویند، μ_w نام دارد و همانند μ_t، برابر است با نسبت مجموع وزن پیوندهای بین انجمن‌ها به مجموع وزن کل پیوندها.

- پارامترهای دیگری نیز وجود دارند مانند حداقل و حداکثر اندازه انجمن‌ها، ضریب خوشه‌بندی میانگین و غیره

این شبکه‌ها ابتدا در سال ۲۰۰۸ برای حالت بدون وزن معرفی شدند [۴۵]. سپس در سال ۲۰۰۹، لانچیکینتی و همکاران در مقالهٔ دیگری حالت وزن‌دار این شبکه‌ها را نیز معرفی کردند [۴۶]. همان‌طور که در این مقاله به تفصیل توضیح داده شده‌است، برای ساخت شبکه‌های LFR وزن‌دار، ابتدا یک شبکه LFR بدون وزن با استفاده از μ_t ساخته می‌شود. سپس وزن‌ها به پیوندها طوری تخصیص داده می‌شوند که $s_i^{(internal)} = (1 - \mu_w)s_i$ باشد. در این معادله s_i قدرت گره i است که به معنی مجموع وزن پیوندهاییست که به گره i متصلند و $s_i^{(internal)}$ به معنی مجموع وزن پیوندهاییست که گره i با گره‌های هم‌انجمن خود برقرار کرده است. توضیحات بیشتر برای نحوهٔ انجام این کار در [۴۶] آمده است. برای تولید شبکه‌های استفاده شده در این پژوهش، از نرم‌افزاری که توسط شخص لانچیکینتی آماده شده و بر روی وبگاه وی قرار گرفته است [۴۷]، استفاده شده‌است.

[۶] mixing parameter for the topology

[۷] mixing parameter for the weights

۴-۳- معیارهای ارزیابی

دو معیار استاندارد برای ارزیابی کارایی الگوریتم‌های پیش‌بینی پیوند استفاده می‌شود: دقت[۸] و *AUC*[۹][۴۸]. اساساً الگوریتم‌های پیش‌بینی پیوند، یک لیست مرتب از تمام پیوندهای ناموجود به ما می‌دهند (معادل $U - E'$) یا به بیان دیگر، به هر پیوند ناموجود (فرض کنید $(x, y) \in U - E'$) یک امتیاز نسبت می‌دهد (فرض کنید S_{xy}) که با استفاده از آن، شانس وجود و یا تشکیل پیوند بین آن دو را کمّی کند. معیار AUC، کارایی الگوریتم را با توجه به کل لیست ارزیابی می‌کند، در حالی که معیار دقت فقط روی L پیوند ابتدایی لیست با بالاترین امتیاز تمرکز می‌کند. تعریف دقیق‌تر این دو معیار به صورت زیر است:

معیار AUC: یک لیست مرتب از امتیاز تمام پیوندهای دیده‌نشده (در مجموعهٔ آزمایش) ساخته می‌شود. مقدار AUC برابر است با احتمال این که یک پیوند گم‌شده (یعنی پیوندی که در مجموعهٔ آزمون قرار دارد $l_1 \in E''$) که به صورت تصادفی انتخاب شده، امتیاز بیشتری نسبت پیوند ناموجودی (یعنی پیوندی که در مجموعه یال‌های گراف اصلی وجود ندارد $l_2 \in U - E$) داشته باشد که آن نیز به صوردت تصادفی انتخاب شده است. یک پیاده‌سازی الگوریتمی برای این معیار به این صورت است: ابتدا برای هر پیوند دیده‌نشده، امتیاز گفته شده محاسبه می‌شود (با روش‌هایی که در ادامه بحث خواهیم کرد). در این حالت خوشبختانه نیازی به مرتب کردن لیست که کاری پرهزینه است نداریم. سپس هر بار یک پیوند از مجموعهٔ پیوندهای گم‌شده و یک پیوند از مجموعه پیوندهای ناموجود، هر دو به صورت تصادفی، انتخاب می‌کنیم امتیاز آن‌ها را مقایسه می‌کنیم. اگر از بین n مقایسهٔ مستقل، در n' حالت امتیاز پیوند گم‌شده از پیوند ناموجود بیشتر بود و در n'' حالت امتیازها برابر بود، مقدار AUC برابر خواهد بود با:

$$AUC = \frac{n' + 0.5n''}{n} \qquad (۱-۴)$$

[۸] Precision

[۹] سطح زیر نمودار منحنی ROC یا به انگلیسی *Area Under the receiver operating characteristic Curve* یا به اختصار AUC

اگر تمام امتیازهایی که به پیوندهای دیده‌نشده اختصاص داده‌ایم، متغیرهای تصادفی مستقل با توزیع یکسان[۱۰] باشند، مقدار AUC میبایست حدود 0.5 به دست آید. بنابراین هر چقدر که این مقدار از 0.5 بیشتر شود، نشان‌دهندهٔ این است که چه مقدار الگوریتم ما از شانس مطلق بهتر عمل می‌کند.

دقت: با داشتن رتبه‌بندی پیوندهای دیده‌نشده، معیار دقت برابر است با نسبت پیوندهایی که درست پیش‌بینی شده‌اند به تعداد پیش‌بینی‌های انجام شده. این معیار به صورت زیر تعریف می‌شود:

$$precision = \frac{TP}{TP + FP} \qquad (۲-۴)$$

که در آن، TP[۱۱] تعداد پیش‌بینی‌های درست و FP[۱۲] تعداد پیش‌بینی‌های غلط است. به عبارت دیگر، اگر ما به تعداد N پیوند از ابتدای لیست پیش‌بینی‌ها برداریم که از این تعداد، N_tp پیش‌بینی درست داشته باشیم (یعنی پیوند پیش‌بینی‌شده‌ای که در مجموعهٔ آزمون یا همان E'' وجود داشته باشد)، معیار دقت برابر خواهد بود با N_tp/N که نام دیگر آن دقت n-بهترین[۱۳] است. واضح است که دقت بیشتر یعنی کارایی بهتر الگوریتم.

علاوه بر این می‌توان برای محاسبهٔ دقت، مقدار N یعنی تعداد پیش‌بینی‌ها را ثابت در نظر نگرفت و دقت را بر حسب تعداد پیش‌بینی‌ها در نموداری ترسیم کرد. این نمودار که دقت را بر حسب تعداد پیش‌بینی‌ها نشان می‌دهد، نمودار دقت-در-n[۱۴] یا به اختصار $P@n$ خوانده می‌شود. منحنی‌هایی را که از رسم $P@n$ به ازای هر n به دست می‌آید، می‌توان علاوه بر مقایسهٔ شهودی، با رابطهٔ میانگین دقت[۱۵] با هم مقایسه کرد که این رابطه به شکل زیر است:

$$AveragePresicion = \frac{\sum_{r=1}^{n} (r \times P@r)}{\sum_{r=1}^{n} r} \qquad (۳-۴)$$

[۱۰] Independent and identically distributed (i.i.d.)

[۱۱] true positive

[۱۲] false positive

[۱۳] Top-n precision

[۱۴] precision at n

[۱۵] Average Precision

با استفاده از شکل ۴- ۱ نحوهٔ محاسبهٔ معیارهای دقت و AUC توضیح داده می‌شود. این گراف ساده، پنج گره، هفت پیوند موجود و سه پیوند ناموجود (پیوندهای $(1,2)$، $(1,4)$ و $(3,4)$) دارد. برای به دست آوردن دقت الگوریتم، ما باید تعدادی از پیوندهای موجود را به عنوان مجموعهٔ آزمون جدا کنیم. به عنوان مثال ما $(1,3)$ و $(4,5)$ را به عنوان مجموعهٔ آزمون انتخاب می‌کنیم که با نقطه‌چین در نمودار سمت چپ مشخص شده‌است. بنابراین هر الگوریتم می‌تواند فقط از اطلاعات مجموعهٔ آموزش استفاده کند که در نمودار سمت چپ با خط نشان داده شده‌اند. فرض کنید یک الگوریتم فرضی به تمام پیوندهای دیده نشده این امتیازها را بدهد: $s_{12} = 0.4, s_{13} = 0.5, s_{14} = 0.6, s_{34} = 0.5, s_{45} = 0.6$. برای محاسبهٔ AUC ما می‌بایست امتیاز پیوندهای دیده نشده و ناموجود را مقایسه کنیم. در کل شش مقایسه وجود دارد: $s_{13} > s_{12}, s_{13} < s_{14}, s_{13} = s_{34}, s_{45} > s_{12}, s_{45} = s_{14}, s_{45} > s_{34}$. بنابراین مقدار AUC برابر خواهد بود با $(3 \times 1 + 2 \times 0.5)/6 \approx 0.67$. برای محاسبهٔ دقت، اگه $L = 2$ در نظر بگیریم، پیوندهای پیش‌بینی‌شده، $(1,4)$ و $(4,5)$ خواهند بود و به دلیل این که یکی درست و یکی غلط است، دقت پیش‌بینی برابر ۰٫۵ خواهد بود.

(آ) شبکه اصلی (ب) شبکه تقسیم شده به دو مجموعهٔ آموزش و آزمون

شکل ۴- ۱: یک شبکهٔ ساده برای توضیح معیارهای ارزیابی

۴-۴- نتایج

در این قسمت، چهار روش پیشنهادی روی چهار شاخص پیش‌بینی پیوند که از دستهٔ معیارهای شباهت مبتنی بر همبندی هستند اعمال می‌شوند و مورد ارزیابی قرار می‌گیرند. این شاخص‌ها عبارتند از:

۱. شاخص همسایگان مشترک (CN)
۲. شاخص آدامیک/ادار (AA)

۳. شاخص تخصیص منابع (RA)
۴. شاخص وابستگی ترجیحی (PA)

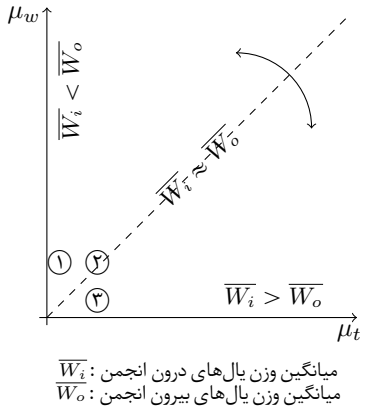
دلیل انتخاب سه شاخص اول، عملکرد بسیار بهتر آن‌ها به نسبت شاخص‌های دیگر (برای مثال شاخص سورنسن یا جاکارد) است. در آزمایشات انجام گرفته، این سه شاخص عملکر عمومی بسیار بهتری از خود نشان می‌دهند و دقت پیش‌بینی آن‌ها زیاد است اما سایر شاخص‌ها دقت مناسبی ندارند. همچنین معیار چهارم یعنی وابستگی ترجیحی نیز به دلیل تفاوت بنیادی آن از نظر ساختاری با سایر شاخص‌ها انتخاب شده است. شاخص‌های دیگر از همه از نظر ساختاری از همسایه‌های مشترک دو گره با ضرایب متفاوت و نرمال‌سازی‌های متفاوت استفاده می‌کنند، اما این شاخص فقط به درجهٔ هر کدام از دو گره توجه می‌کند و به خاطر همین موضوع، پیچیدگی زمانی آن نیز با سایر شاخص‌ها متفاوت است و در حالی که سایر شاخص‌ها پیچیدگی زمانی از مرتبه $O(n^2)$ (یا ضرایب آن) دارند، این شاخص از پیچیدگی زمانی $O(n)$ برخوردار است.

همان‌طور که پیش‌تر گفته شد، برای روش تشخیص انجمن نیز از روش نقشه‌اطلاعات استفاده می‌شود. این انتخاب روش بر پایهٔ مطالعهٔ جامع لانچیکینتی و فورتوناتو[۱۶] [۴۹] انجام گرفته است. در این مطالعه، این دو محقق نشان دادند که الگوریتم نقشه‌اطلاعات به طور کلی روی شبکه‌های LFR بهترین عملکرد را دارد و در کل نیز یکی از بهترین روش‌های تشخیص انجمن به شمار می‌رود.

پیش‌تر اشاره شد که برای انجام آزمایشات، از روش ارزیابی متقاطع ۱۰-قسمتی استفاده خواهد شد و نتایج گزارش شده، میانگین ۱۰ بار اجرای الگوریتم خواهد بود. البته در بعضی آزمایش‌ها برای بالا بردن تعداد تکرارها، بیش از یک بار (برای مثال ۳ یا ۵ بار) از ارزیابی متقاطع ۱۰-قسمتی استفاده شده‌است.

هدف این پژوهش در بخش‌های گذشته بررسی عملکرد روش‌های پیشنهادی روی پارامترهای شبکه‌های LFR عنوان شد. دو پارامتر اصلی که برای ما از اهمیت ویژه برخوردارند و نحوه توزیع پیوندها و وزن آن‌ها را این شبکه‌ها کنترل می‌کنند، دو پارامتر μ_t و μ_w هستند. در نتیجه برای بررسی اثر این پارامتر می‌بایست با تغییر این پارامترها،

[۱۶] Fortunato



شکل ۴- ۲: نحوه توزیع وزن‌ها در محیط پارامتری دو بعدی $\mu_t - \mu_w$

آزمایشات لازم را انجام داده و نتایج آن‌ها با هم مقایسه شوند. برای این کار مقادیر هر یک از این پارامترها را در بازهٔ $[0.1, 0.7]$ با گام 0.1 تغییر داده و آزمایشات لازم را انجام می‌شوند. این کار باعث می‌شود که یک فضای پارامتر دو بعدی متشکل از $7 \times 7 = 49$ نقطه داشته باشیم. در ادامه، نتایج آزمایش‌ها برای این ۴۹ نقطه در فضای پارامتر دو بعدی بررسی خواهند شد.

ما این فضای پارامتر دو بعدی را به سه منطقه مانند شکل ۴- ۲ تقسیم می‌کنیم: ۱) منطقه‌ای که μ_t از μ_w کوچکتر است؛ ۲) منطقه‌ای که μ_t با μ_w (به صورت تقریبی) برابر است؛ ۳) منطقه‌ای که μ_t از μ_w بزرگتر است. با توجه به تعریف μ_t و μ_w، نسبت μ_t از یال‌های شبکه درون انجمن‌ها قرار می‌گیرند و به همین ترتیب $1 - \mu_w$ از مجموع وزن یال‌ها درون انجمن‌ها قرار می‌گیرند. این به این معناست که اگر $\mu_t < \mu_w$ باشد، یعنی در منطقه ۱، یال‌هایی که بیرون از انجمن‌ها قرار می‌گیرند به طور میانگین وزن بیشتری نسبت به یال‌هایی دارند که درون انجمن‌ها قرار می‌گیرند، چون نسبت تعداد آن‌ها بیشتر است اما نسبت وزن آن‌ها به آن اندازه نیست. به طور مشابه، وقتی $\mu_t = \mu_w$ باشد، یعنی در منطقه ۲، میانگین وزن یال‌های درون انجمن با یال‌های بیرون انجمن برابر است و وقتی $\mu_t > \mu_w$ باشد، یعنی در منطقه ۳، میانگین وزن یال‌های درون انجمن بیشتر از میانگین وزن یال‌های بیرون انجمن است. اهمیت این موضوع در این است که ما می‌خواهیم پیش‌بینی پیوند را به درون انجمن‌ها محدود کنیم و اثر بخش‌های مختلف فضای پارامتری را بر روی این کار بررسی کنیم.

جدول ۴- ۱: پارامترهای مورد استفاده برای شبکه‌های LFR مورد استفاده در آزمایشات

تعداد گره‌ها	۵۰۰
میانگین درجهٔ گره‌ها	۱۰
بیشینهٔ درجهٔ گره‌ها	۵۰
کمینهٔ اندازهٔ انجمن‌ها	۴
بیشینهٔ اندازهٔ انجمن‌ها	۱۰۰

برای سایر پارامترهای شبکه، مقادیر مختلفی آزمایش شده‌است. نتایج کلی به دست آمده در این پژوهش، نسبت به سایر پارامترها الگوی یکسانی را از خود نشان می‌دهد و تقریباً ثابت است. برای آزمایش‌های گزارش شده در این پژوهش، از مقادیر موجود در جدول ۴- ۱ استفاده شده‌است.

۴- ۴- ۱ نتایج حاصل از با ارزیابی معیار دقت n-بهترین

در این بخش مطابق معمول آزمایش‌های پیش‌بینی پیوند، مقدار n را برابر با ۱۰۰ در نظر می‌گیریم. به دلیل این که تعداد حالت‌های مختلف که می‌بایست این معیار دقت n-بهترین را در آن‌ها گزارش کنیم بسیار زیاد است (چهار شاخص، هر کدام در ترکیب با روی چهار روش پیشنهادی، روی ۴۹ نقطه فضای پارامتری) بخشی از این نتایج را به عنوان نمونه گزارش می‌کنیم. الگوی کلی در نتایج نمونه با الگوی کلی در تمام نتایج یکسان است.

در جدول ۴- ۲ نتایج دقت n-بهترین برای شاخص AA به همراه چهار روش پیشنهادی در حالتی که پارامتر μ_t ثابت و برابر با 0.3 است و پارامتر μ_w بین 0.1 و 0.7 تغییر می‌کند، نشان داده شده‌است. در این جدول، اعدادی که پررنگ نشان داده شده‌اند دو روش اولی هستند که نتیجه‌ای بهتر از روش پایه داشته‌اند.

جدول ۴- ۲: نتایج به دست آمده برای معیار AA همراه چهار روش پیشنهادی در $\mu_t = 0.3$

	AA	$AA+WW$	$AA+WU$	$AA+UU$	$AA+UW$
$\mu_w = 0.1$	0.212(35)	**0.213(35)**	**0.213(35)**	0.203(40)	0.203(40)
$\mu_w = 0.2$	0.236(44)	**0.240(45)**	**0.240(45)**	0.234(41)	0.234(41)
$\mu_w = 0.3$	0.189(30)	0.210(29)	0.210(29)	**0.211(32)**	**0.211(32)**
$\mu_w = 0.4$	0.140(26)	0.174(27)	0.176(28)	**0.184(32)**	**0.182(31)**
$\mu_w = 0.5$	0.100(23)	0.048(26)	**0.183(28)**	**0.192(32)**	0.066(32)
$\mu_w = 0.6$	0.091(28)	0.007(8)	**0.215(42)**	**0.240(43)**	0.015(14)
$\mu_w = 0.7$	0.035(17)	0.015(13)	**0.137(31)**	**0.168(37)**	0.017(14)

جدول ۴- ۳ نیز متقابلاً دقت n-بهترین را برای روش‌های یاد شده در حالتی نشان می‌دهد که μ_t متغیر است و μ_w ثابت در نظر گرفته شده است.

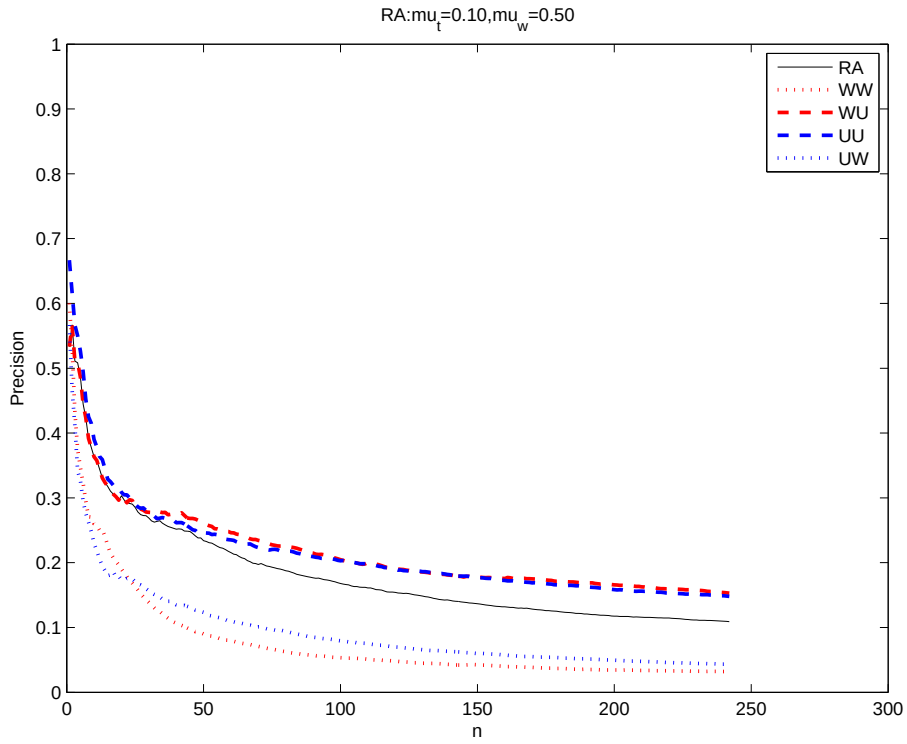
جدول ۴- ۳: نتایج به دست آمده برای معیار AA همراه چهار روش پیشنهادی در $\mu_w = 0.3$

	AA	$AA+WW$	$AA+WU$	$AA+UU$	$AA+UW$
$\mu_t = 0.1$	0.194(39)	0.174(55)	**0.233(41)**	**0.240(45)**	0.222(37)
$\mu_t = 0.2$	0.305(41)	**0.334(41)**	**0.334(41)**	0.323(41)	0.323(41)
$\mu_t = 0.3$	0.186(30)	**0.211(31)**	**0.211(31)**	0.195(34)	0.195(34)
$\mu_t = 0.4$	0.118(24)	**0.137(19)**	**0.137(18)**	0.123(26)	0.123(26)
$\mu_t = 0.5$	0.078(27)	**0.083(30)**	**0.080(27)**	0.065(25)	0.073(27)
$\mu_t = 0.6$	0.076(23)	**0.090(21)**	0.066(27)	0.051(18)	**0.086(26)**
$\mu_t = 0.7$	0.031(15)	0.030(15)	0.022(13)	0.015(12)	**0.032(14)**

اما برای این که بتوان نتیجه‌گیری بهتری از نتایج داشت نیاز است تا معیارهایی داشته باشیم که بازنمایی بهتری داشته باشند. یکی از چالش‌های معیار دقت n-بهترین، تعیین n است که می‌تواند تاثیر گذار باشد. به همین دلایل، استفاده از معیاری مثل معیار دقت در n می‌تواند نتایج را روشن‌تر به ما نشان دهد.

۴- ۴- ۲ نتایج حاصل از ارزیابی با معیار دقت در n

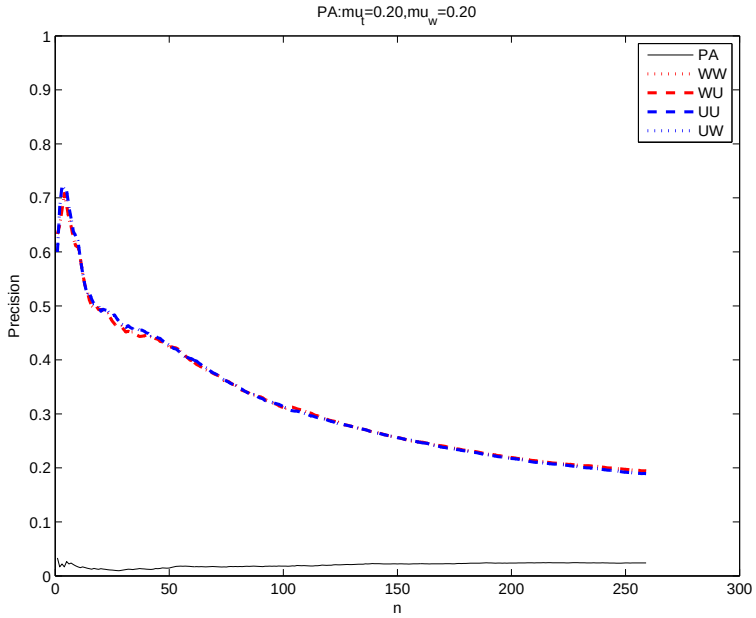
همان‌طور که در بخش ۴- ۳- عنوان شد، این معیار، دقت را به ازای nهای مختلف در یک نمودار نشان می‌دهد که کمک می‌کند که بتوان مقایسه‌ای به صورت شهودی بین روش‌ها انجام داد. همانند بخش گذشته، به دلیل این که تعداد این نمودارها زیاد است، مجبوریم به ارائهٔ نمونه‌ای از آن‌ها بسنده کنیم. هر نمودار مقایسهٔ روش پایه (یکی از چهار شاخص شباهت انتخاب شده) با چهار روش پیشنهادی که با همان شاخص ترکیب شده‌اند را در یکی از نقاط فضای پارامتر نشان می‌دهد. برای نمونه، شکل ۴- ۳ منحنی‌های دقت در n روش RA را به همراه چهار روش پیشنهادی که روی روش RA اعمال شده‌اند، در نقطهٔ $\mu_t = 0.1, \mu_w = 0.5$ نشان می‌دهد. در این شکل واضح است که در این نقطه از فضای پارامترها، روش‌های WU و UU بسیار نزدیک به هم هستند و از روش پایهٔ RA کارایی بهتری دارند و دقت بیشتری ارائه می‌دهند، اما از سوی دیگر روش‌های WW و UW دقت روش پایه را کاهش می‌دهند. در ادامه، دلیل این رفتار مورد بررسی قرار خواهد گرفت.



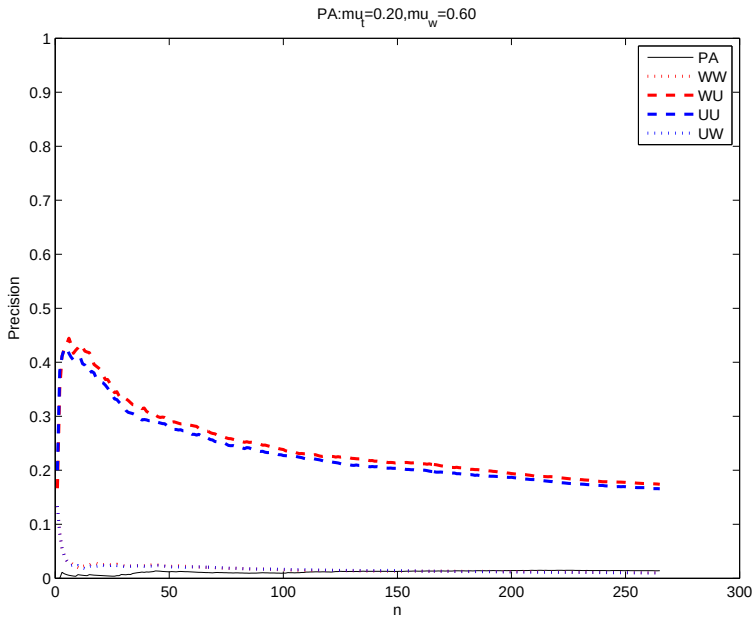
شکل ۴- ۳: معیار دقت در n برای شاخص RA و روش‌های پیشنهادی در نقطهٔ $\mu_t = 0.1, \mu_w = 0.5$

شکل‌های ۴- ۴ تا ۴- ۷ نیز شاخص PA را به همراه چهار روش پیشنهادی در چهار نقطه مختلف از فضای پارامتری نشان می‌دهند. همان‌طور که در شکل ۴- ۴ مشاهده می‌شود، در نقطهٔ $\mu_t = 0.2, \mu_w = 0.2$ (که در منطقهٔ ۲ قرار دارد) با استفاده از هر چهار روش پیشنهادی می‌توان به اندازهٔ قابل ملاحظه‌ای دقت پیش‌بینی را بهبود بخشید. در شکل ۴- ۵ مشاهده می‌شود که در نقطهٔ $\mu_t = 0.2, \mu_w = 0.6$ (که در منطقهٔ ۱ قرار دارد) روش‌های UU و WU توانسته‌اند به بهبود دقت پیش‌بینی کمک شایانی بکنند اما دو روش دیگر یعنی WW و UW بهبودی صورت نداده‌اند. در شکل بعد یعنی شکل ۴- ۶ در نقطهٔ $\mu_t = 0.6, \mu_w = 0.2$ (که در منطقهٔ ۳ قرار دارد) برعکس شکل قبل روش‌های WW و UW خوب عمل کرده‌اند اما روش‌های UU و WU نتوانسته‌اند کمک چندانی به پیش‌بینی کنند. در نهایت در شکل ۴- ۷ در نقطهٔ $\mu_t = 0.6, \mu_w = 0.6$ (که باز هم در منطقهٔ ۲ قرار دارد) هر چهار معیار اندکی توانسته‌اند به بهبود پیش‌بینی کمک کنند.

اما برای مشاهده تمام این نقاط با هم در یک نمودار واحد می‌بایست روش دیگری را به کار گیریم. در بخش بعد روشی برای این کار معرفی می‌کنیم.



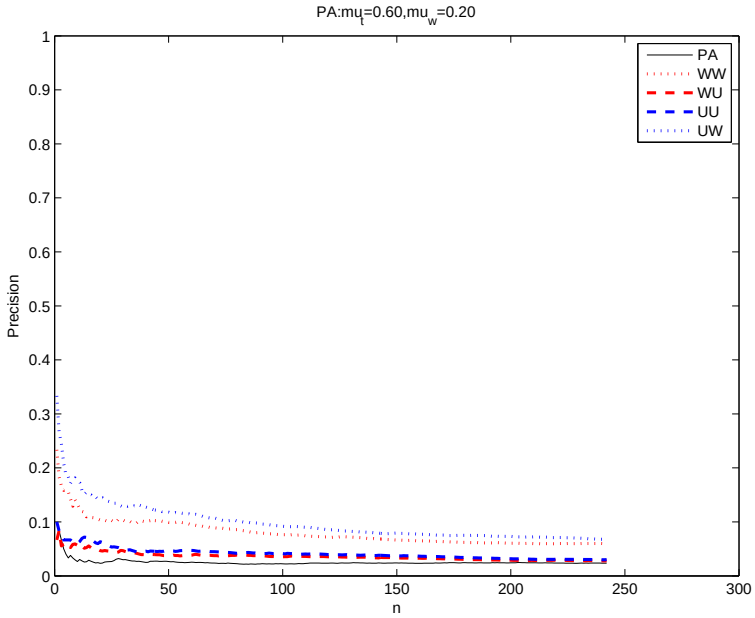
شکل ۴- ۴: معیار دقت در n برای شاخص PA و روش‌های پیشنهادی در نقطهٔ $\mu_t = 0.2, \mu_w = 0.2$



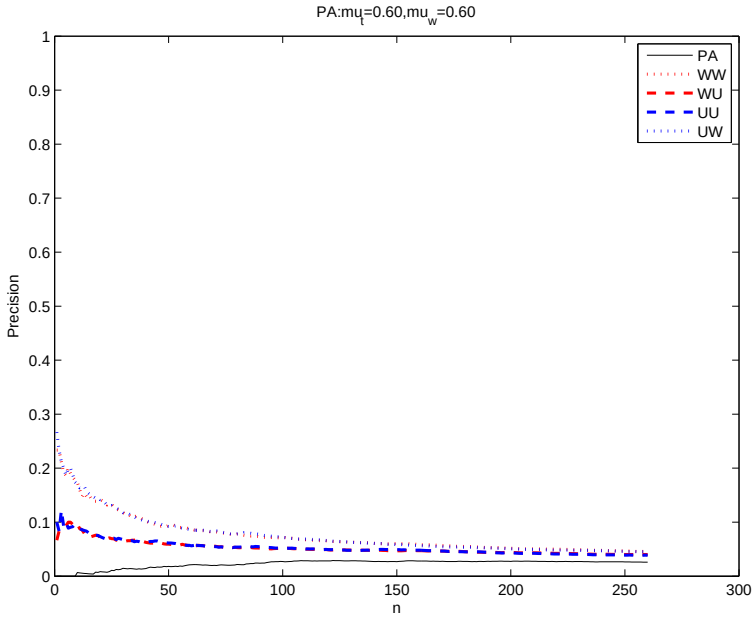
شکل ۴- ۵: معیار دقت در n برای شاخص PA و روش‌های پیشنهادی در نقطهٔ $\mu_t = 0.2, \mu_w = 0.6$

۴- ۴- ۳ نتایج حاصل از ارزیابی با معیار دقت میانگین

همان‌طور که در بخش‌های قبل اشاره کردیم، هر کدام از منحنی‌های نمودارهای دقت در n را می‌توان با استفاده از رابطهٔ ۴- ۳ به یک عدد تبدیل کرد که دقت متوسط نام دارد. از این معیار برای مشاهدهٔ کارایی روش‌ها در نقاط مختلف



شکل ۴- ۶: معیار دقت در n برای شاخص PA و روش‌های پیشنهادی در نقطهٔ $\mu_t = 0.6, \mu_w = 0.2$



شکل ۴- ۷: معیار دقت در n برای شاخص PA و روش‌های پیشنهادی در نقطهٔ $\mu_t = 0.6, \mu_w = 0.6$

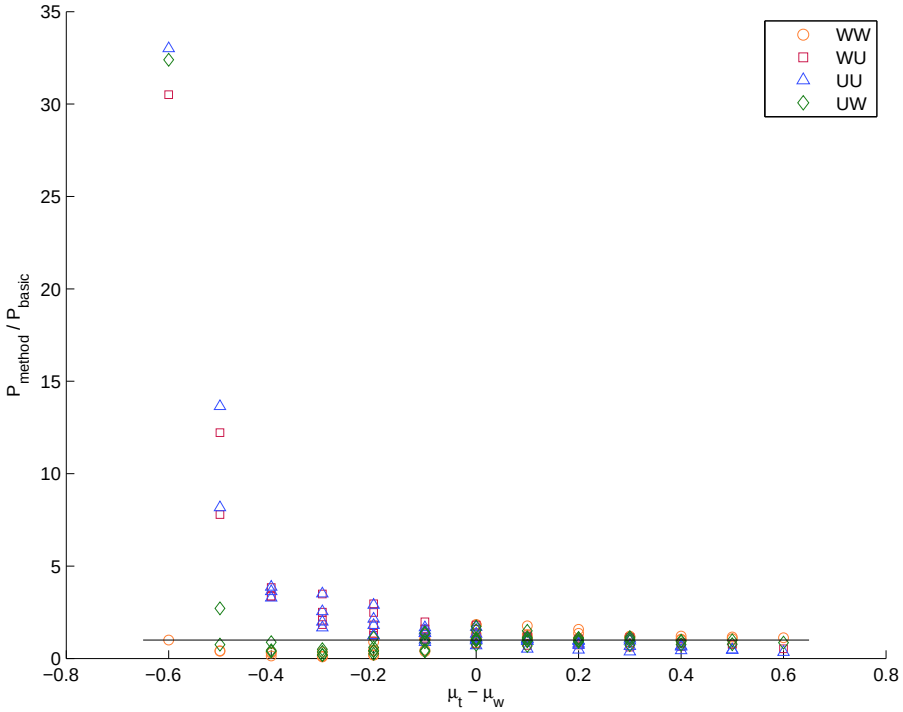
فضای پارامتری در کنار هم استفاده می‌کنیم. برای این کار، هر کدام از منحنی‌هایی که در نمودارهای دقت در n وجود دارند، به یک عدد تبدیل می‌کنیم، سپس نسبت و یا اختلاف هر یک از چهار روش پیشنهادی در مقایسه با روش پایه به دست می‌آوریم. در نهایت این نسبت و یا را در نموداری رسم می‌کنیم.

با توجه به این که فضای پارامتر ما دو بعدی است و در نتیجه رسم این مقادیر نیاز به فضای سه‌بعدی دارد (که ارائه آن در گزارش امکان‌پذیر نیست)، فضای پارامتر دو بعدی را با استفاده از تبدیل $\mu_t - \mu_w$ به فضای یک بعدی تبدیل می‌کنیم. در نتیجه می‌توان مقادیر به دست آمده را در یک نمودار دو بعدی نشان داد که محور افقی $\mu_t - \mu_w$ و محور عمودی نسبت یا اختلاف را نشان می‌دهد. با استفاده از این تبدیل، می‌توان اثر سه منطقهٔ بحث شده را نیز مشاهده کرد به این صورت که سمت چپ نمودار تا مقدار صفر معادل منطقهٔ ۱، وسط نمودار معادل منطقهٔ ۲ و سمت راست نمودار معادل منطقهٔ ۳ خواهد بود.

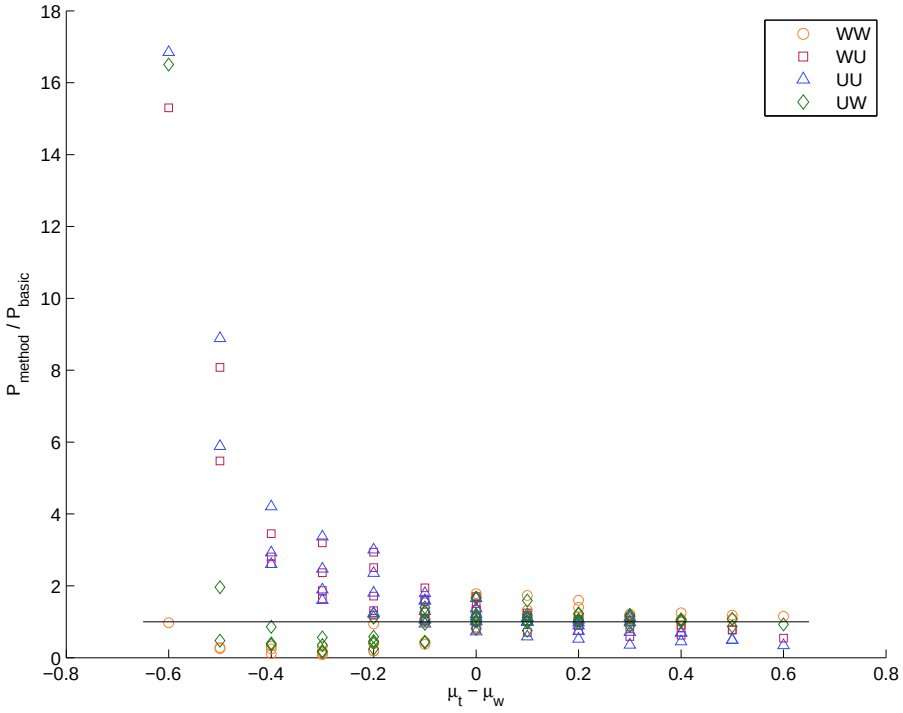
شکل‌های ۴- ۸ تا ۴- ۱۱ این نمودارها را به ترتیب برای شاخص‌های CN، AA، RA و PA در حالتی که نسبت دقت پیش‌بینی مورد نظر بوده، نشان می‌دهند. هر کدام از نقاط روی این شکل‌ها نسبت دقت پیش‌بینی هر یک از روش‌های پیشنهادی به دقت پیش‌بینی روش پایه است که بر حسب اختلاف دو پارامتر μ_t و μ_w رسم شده‌اند. همان‌طور که در این شکل‌ها مشخص است در منطقهٔ یک که سمت چپ نمودار است، استفاده از روش‌های WU و UU توانسته دقت پیش‌بینی بهتری از حالت پایه ارائه کند، اما دو روش دیگر یعنی WW و UW نتوانسته‌اند تاثیر مثبتی روی دقت پیش‌بینی داشته باشند. هر چه از سمت چپ نمودار به سمت راست حرکت کنیم که معادل حرکت از منطقهٔ ۱ به سمت منطقهٔ ۲ و در نهایت منطقهٔ ۳ است، روش‌های WU و UU افت کارایی زیادی دارند و از طرف دیگر کارایی دو روش دیگر یعنی WW و UW بهتر می‌شود.

شکل‌های ۴- ۱۲ تا ۴- ۱۵ نیز مشابه شکل‌های شکل‌های ۴- ۸ تا ۴- ۱۱ هستند با این تفاوت که به جای نسبت، اختلاف دقت پیش‌بینی هر یک از روش‌های پیشنهادی با دقت پیش‌بینی روش پایه را نشان می‌دهند.

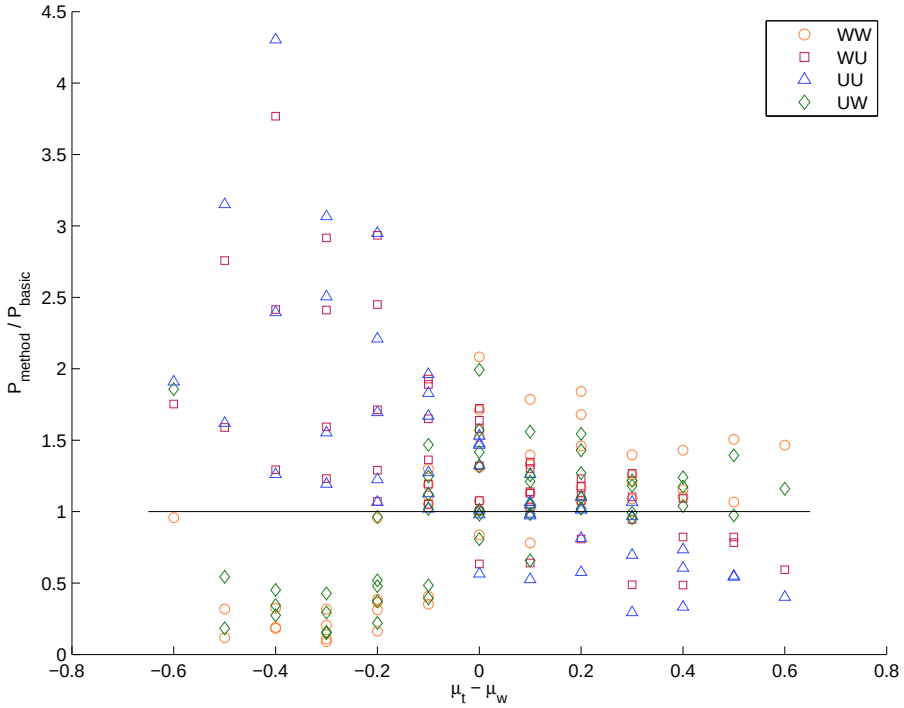
در تحلیل رفتار این روش‌ها در نمودارها، می‌توان گفت که دلیل اصلی این است که در منطقه ۱، یال‌های بین انجمن‌ها قوی‌تر از یال‌های درون انجمن‌هاست. همین قوی‌تر بودن یای‌های بین انجمن‌ها باعث می‌شود که روش‌های تشخیص انجمن وزن‌دار به اشتباه بیافتند و نتواند به نحو مناسبی انجمن‌ها را تشخیص دهند. همین امر باعث می‌شود که دو روش پیشنهادی که از تشخیص انجمن وزن‌دار استفاده می‌کنند، کارایی کمتری داشته باشند و حتا کارایی آن‌ها از روش پایه نیز بدتر شود. اما روش‌های تشخیص انجمن بدون وزن، به دلیل در نظر نگرفتن وزن یال‌ها، خللی در



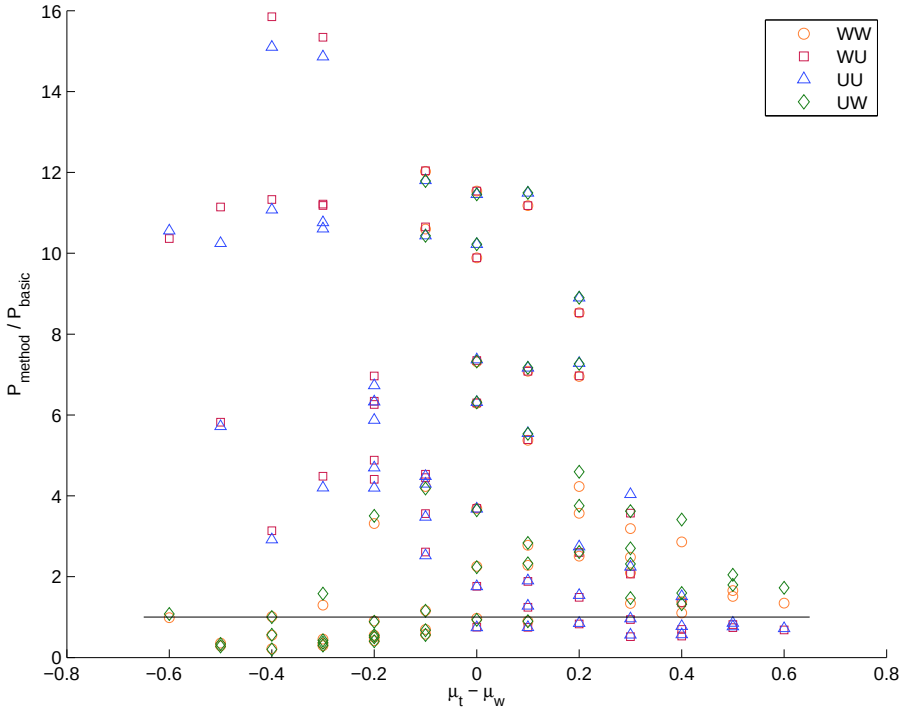
شکل ۴- ۸: نسبت بهبود کارایی روش‌های پیشنهادی در معیار CN



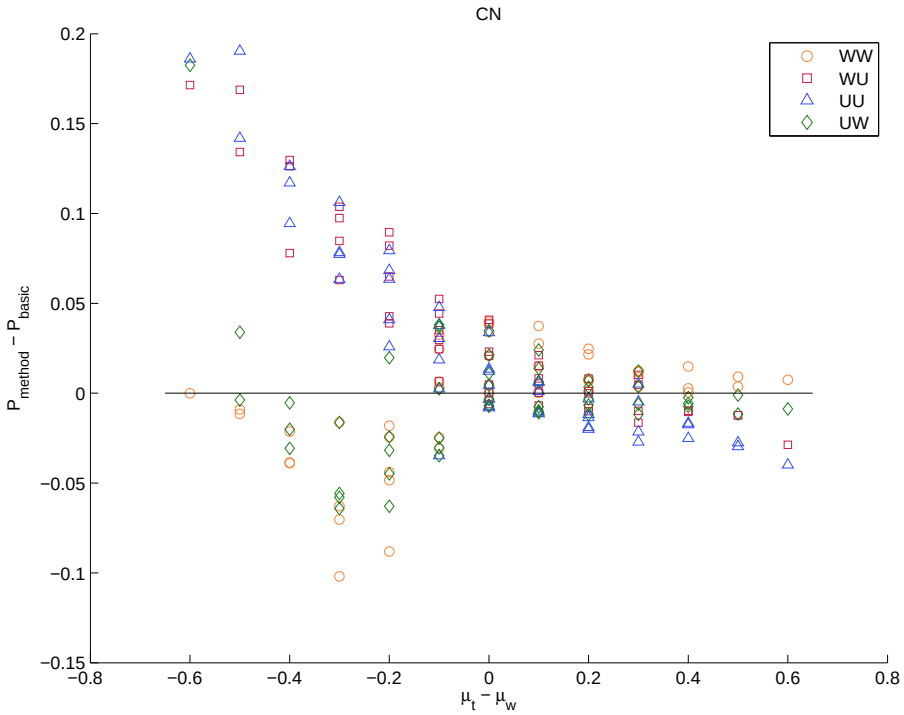
شکل ۴- ۹: نسبت بهبود کارایی روش‌های پیشنهادی در معیار AA



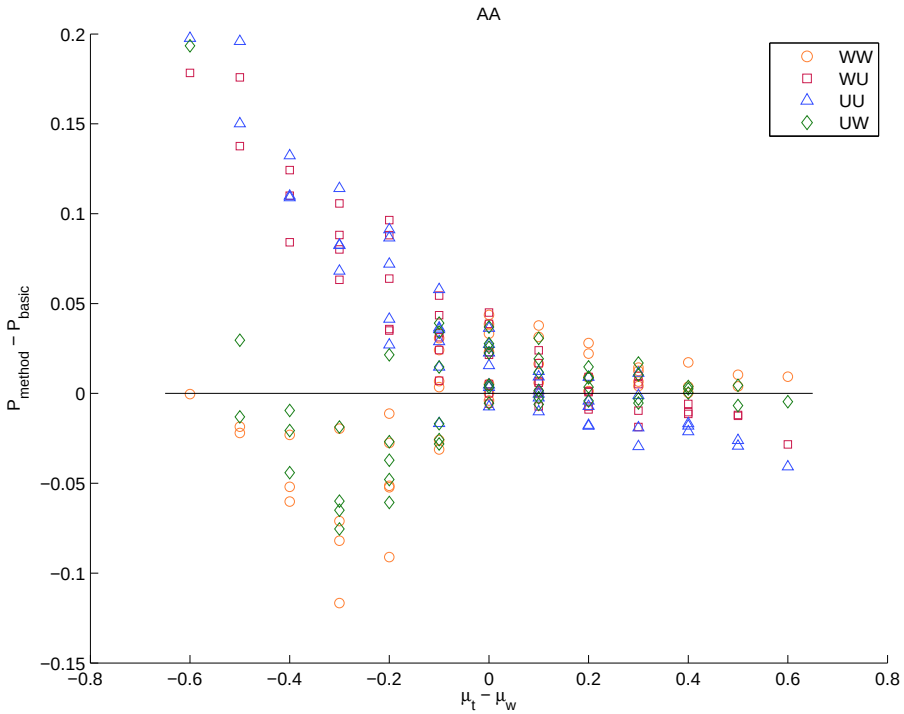
شکل ۴- ۱۰: نسبت بهبود کارایی روش‌های پیشنهادی در معیار RA



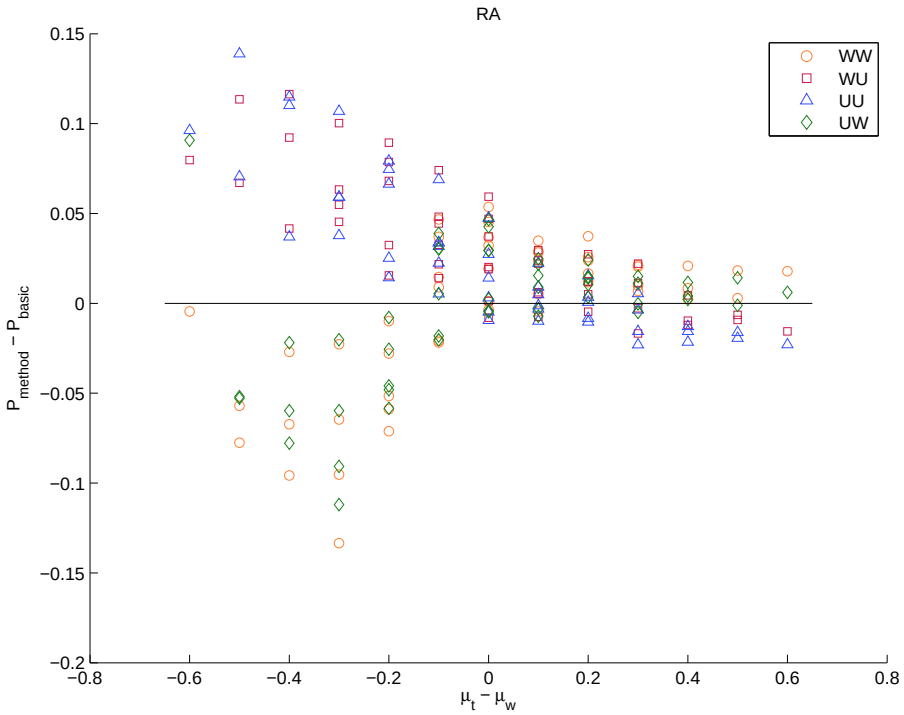
شکل ۴- ۱۱: نسبت بهبود کارایی روش‌های پیشنهادی در معیار PA



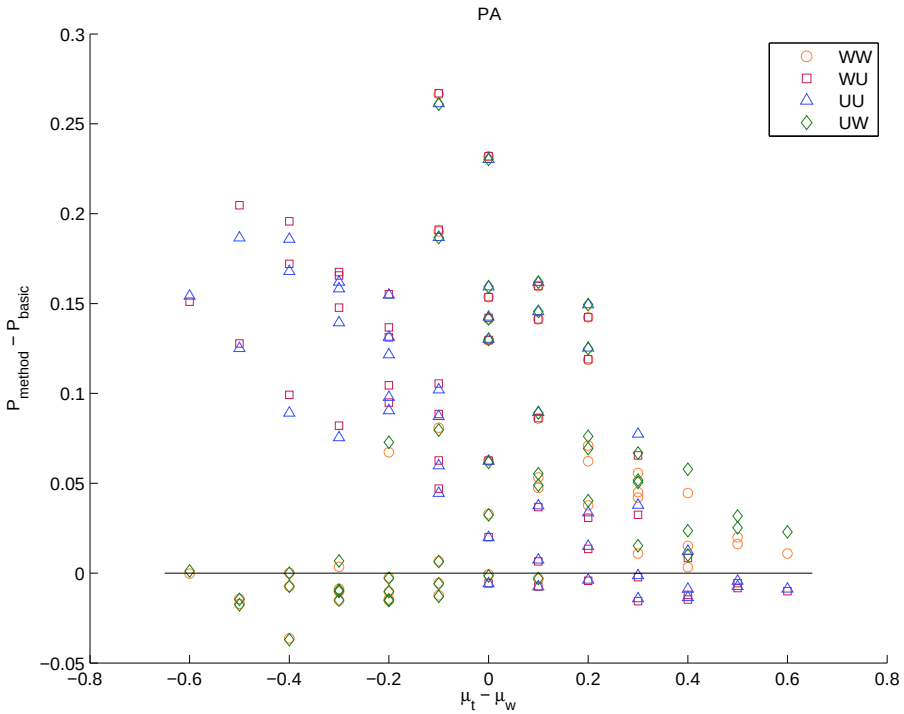
شکل ۴- ۱۲: میزان بهبود کارایی روش‌های پیشنهادی در معیار CN



شکل ۴- ۱۳: میزان بهبود کارایی روش‌های پیشنهادی در معیار AA



شکل ۴- ۱۴: میزان بهبود کارایی روش‌های پیشنهادی در معیار RA



شکل ۴- ۱۵: میزان بهبود کارایی روش‌های پیشنهادی در معیار PA

کارشان ایجاد نمی‌شود و می‌توانند به خوبی انجمن‌ها را تشخیص دهند و پیش‌بینی روش‌های پایه را بهبود ببخشند.

اما هر چه از منطقهٔ ۱ فاصله می‌گیریم و به سمت منطقهٔ ۳ می‌رویم، وزن‌یال‌ها درون انجمن‌ها قوی‌تر می‌شوند و در نهایت قوی‌تر از یال‌های بیرون انجمن می‌شوند. این موضوع باعث می‌شود که روش‌های تشخیص انجمنی که از وزن استفاده می‌کنند، انجمن‌های بهتری تشخیص دهند و در نتیجه با این تشخیص انجمن مناسب بتوانند به دقت‌های بهتری برسند. در مقابل نیز روش‌هایی که از اطلاعات وزن استفاده نمی‌کنند با افت کیفیت مواجه می‌شوند چون وزن در این منطقه یک عامل کلیدی است.

۴-۴-۴ نتایج حاصل از ارزیابی با معیار AUC

در این بخش نتایج ارزیابی با معیار AUC بررسی می‌شود. همانند بخش‌های گذشته به علت حجم بالای خروجی به ارائه نمونه‌ای از آن بسنده می‌شود. سایر نتایج از الگوی مشابهی پیروی می‌کنند. جدول ۴-۴ مقدار AUC را به ازای $\mu_w = 0.3$ و μ_t متغیر و جدول ۴-۵ نیز مقدار AUC را به ازای $\mu_t = 0.3$ و μ_w متغیر نشان می‌دهد. همان‌طور که از این جداول برمی‌آید، روش‌های پیشنهادی در بیشتر موارد نتوانسته‌اند مقدار AUC را بهبود ببخشند. دلیل این امر این است که روش‌های پیشنهادی، با توجه به ساختارشان نمی‌توانند یال‌هایی که بین انجمن‌ها را تشخیص دهند و همین باعث می‌شود که مقدار AUC افت پیدا کند. البته از طرف دیگر پیش‌بینی‌های اشتباه بین انجمن‌ها نیز رخ نمی‌دهد و این ممکن است باعث بالا رفتن مقدار AUC شود. می‌توان گفت برآیند این دو موضوع در حالت کلی تغییر چندانی در مقدار AUC ایجاد نمی‌کند و همان‌طور که از جدول‌ها مشخص است در اکثر موارد مقادیر AUC به هم نزدیکند. البته باید توجه داشت که معیار AUC معیاریست که کل پیش‌بینی‌ها برای محاسبهٔ آن در نظر گرفته می‌شود. در حالی که معمولاً در کاربردهای مسئلهٔ پیش‌بینی پیوند، کارایی پیش‌بینی در پیش‌بینی‌های ابتدایی مهمتر است تا پیش‌بینی‌های انتهایی. در نتیجه معیارهای دقت اهمیت بیشتری پیدا می‌کنند.

جدول ۴- ۴: نتایج حاصل از ارزیابی معیار AUC برای شاخص CN به ازای $\mu_w = 0.3$

	CN	$CN+WW$	$CN+WU$	$CN+UU$	$CN+UW$
$\mu_t = 0.1$	0.876	0.875	0.877	0.874	0.872
$\mu_t = 0.2$	0.776	0.778	0.774	0.773	0.775
$\mu_t = 0.3$	0.781	0.774	0.763	0.763	0.773
$\mu_t = 0.4$	0.702	0.681	0.657	0.657	0.681
$\mu_t = 0.5$	0.679	0.645	0.601	0.601	0.645
$\mu_t = 0.6$	0.652	0.601	0.558	0.557	0.599
$\mu_t = 0.7$	0.640	0.549	0.602	0.601	0.549

جدول ۴- ۵: نتایج حاصل از ارزیابی معیار AUC برای شاخص CN به ازای $\mu_t = 0.3$

	CN	$CN+WW$	$CN+WU$	$CN+UU$	$CN+UW$
$\mu_w = 0.1$	0.781	0.774	0.763	0.763	0.773
$\mu_w = 0.2$	0.755	0.742	0.740	0.738	0.742
$\mu_w = 0.3$	0.754	0.728	0.738	0.737	0.726
$\mu_w = 0.4$	0.743	0.660	0.738	0.736	0.660
$\mu_w = 0.5$	0.744	0.508	0.744	0.742	0.508
$\mu_w = 0.6$	0.753	0.505	0.759	0.757	0.505
$\mu_w = 0.7$	0.762	0.510	0.769	0.768	0.511

۴- ۵- جمع‌بندی

در این فصل، پس از بیان مقدمه، به معرفی مجموعه داده‌های مورد استفاده در پژوهش حاضر یعنی شبکه‌های LFR پرداخته شد و بیان شد که این شبکه‌ها نوعی شبکهٔ ساختگی از خانواده شبکه‌های مقیاس آزاد هستند و در مورد پارامترهای آن‌ها نیز صحبت شد. سپس به توضیح در مورد معیارهای ارزیابی روش‌های پیش‌بینی پیوند پرداخته شد. پس از آن روش‌های پیشنهادی روی روی فضای پارامتری شبکه‌های LFR اعمال شد و نتایج آن‌ها با توجه به معیارهای ارزیابی مورد بحث، عنوان شد. در نهایت نتایج به دست آمده تحلیل شد. در تحلیل نتایج گفته شد که هر کدام از روش‌های پیشنهادی، در کدام قسمت از فضای پارامتر شبکه‌های LFR کارایی خوبی دارند و در کدام قسمت کارایی خوبی از خود نشان نمی‌دهند. دلایل این اختلاف کارایی در قسمت‌های مختلف نیز مورد تحلیل و بررسی قرار گرفت.

فصل ۵: بحث و نتیجه‌گیری

۵- ۱- مقدمه

در این فصل از پایان‌نامه، به بحث و نتیجه‌گیری از تلاش‌های صورت گرفته در این پژوهش پرداخته خواهد شد. ابتدا خلاصه‌ای از هدف تحقیق بیان می‌شود. سپس روش‌های پیشنهادی و تفاوت آن‌ها با کارهای پیشین در این حوزه بیان می‌شود. در ادامه به اختصار دربارهٔ چگونگی انجام پژوهش و نتایج به دست آمده از آن بحث می‌شود. در آخر نیز کارهای آینده و پیشنهادات مطرح خواهند شد.

۵- ۲- جمع‌بندی

هدف اصلی از انجام این پژوهش، کمک به بهبود روش‌های پیش‌بینی پیوند موجود بوده است. همان‌طور که عنوان شد، پیش‌بینی پیوند یک مسئلهٔ اساسی و مهم در دانش اطلاعات مدرن است که در حوزه‌های مختلف از تحلیل

شبکه‌های اجتماعی گرفته تا زیست‌شناسی و غیره کاربرد دارد و همین موضوع، نیاز به روش‌های کارا برای حل این مسئله را روشن می‌کند. بر اساس همین هدف، تلاش‌هایی در همین راستا صورت گرفته است که لیست دستاوردهای آن به شرح زیر است:

- مطالعهٔ کاملی از انواع روش‌های پیش‌بینی پیوند انجام شد که دسته‌بندی و توضیحات مربوط به هر یک از روش‌ها در فصل مروری بر ادبیات آورده شده‌است.

- ایدهٔ اصلی این پژوهش که استفاده از وزن پیوندها به همراه اطلاعات انجمن‌ها بود مطرح شد و به همین منظور پیشینه مختصری از روش‌ها تشخیص انجمن‌ها و روش مورد نظر این پژوهش ارائه شد.

- روش پیشنهادی در این پژوهش که معرفی شد. توضیح داده شد که چرا اطلاعات انجمن‌ها برای پیش‌بینی پیوند مفیدند. روش اصلی پیشنهادی محدود کردن روش‌های پیش‌بینی پیوند به پیش‌بینی در داخل انجمن‌ها بود. دلیل این امر این است که از طرفی معمولاً گره‌ها تمایل بیشتری به ایجاد پیوند با گره‌های هم‌انجمن خود دارند و با احتمال بیشتری به آن‌ها متصل خواهند شد و از طرف دیگر تعداد یال‌های بالقوه درون انجمن‌ها بسیار کمتر از یال‌های بالقوه خارج از انجمن‌هاست و همین دو نکته باعث می‌شود که پیش‌بینی داخل انجمن بتواند کمک خوبی به بهبود کارایی روش‌های پیش‌بینی پیوند بکند.

- روش پیشنهادی به دو گام محاسبهٔ شاخص شباهت و تشخیص انجمن تقسیم شد. توضیح داده شد که استفاده از وزن پیوندها در هر کدام از این گام‌ها امکان‌پذیر است. در نتیجه با توجه به استفاده کردن یا نکردن از وزن پیوندها در هر کدام از این دو گام، روش پیشنهادی به چهار روش تبدیل می‌شود که هر کدام از این چهار روش می‌توانند با توجه به ویژگی‌های شبکه‌ای که در آن قرار است استفاده شوند، کارایی‌های متفاوتی از خود نشان دهند.

- برای انجام آزمایش و آزمودن روش‌های پیشنهادی، یک نوع شبکهٔ مصنوعی پارامتری به نام شبکه‌های LFR انتخاب شد و پیشینه این شبکه‌ها بررسی شد. شبکه‌های LFR نوعی شبکهٔ مستقل از اندازه هستند که بررسی

آن‌ها می‌تواند نکات ارزشمندی در اختیار ما بگذارد. از آن‌جایی که این شبکه‌ها، شبکه‌های پارامتری هستند، بررسی روش‌های پیشنهادی روی فضای پارامتری این شبکه‌ها که متشکل از دو پارامتر μ_t و μ_w است انجام شد. این پارامترها به ترتیب نحوهٔ توزیع یال‌ها در درون و بیرون از انجمن‌ها، و نحوهٔ توزیع وزن یال‌ها باز هم در درون و بیرون از انجمن‌ها را تعیین می‌کنند.

- در نهایت بعد از بررسی آزمایشات انجام شده، نتیجه‌گیری شد که در حالت‌هایی که μ_t از μ_w بیشتر است، یعنی وقتی که یال‌های درون انجمن‌ها قوی‌تر از یال‌های بیرون انجمن‌ها هستند، روش‌هایی که از تشخیص انجمن وزن‌دار استفاده می‌کنند کارایی بهتری از حالت معمولی (بدون استفاده از انجمن‌ها دارند) ولی روش‌هایی که از تشخیص انجمن بدون وزن استفاده می‌کنند کارایی مطلوبی ندارند به این دلیل که وزن زیاد یال‌ها درون انجمن‌ها، به تشخیص انجمن کمک می‌کند. از طرف دیگر در حالت‌هایی که μ_w از μ_t بیشتر است، روش‌هایی که از تشخیص انجمن بدون وزن استفاده می‌کنند بسیار بهتر از حالت معمولی و دو روش دیگر عمل می‌کنند به این دلیل که وزن زیاد یال‌های خارج از انجمن، باعث به خط افتادن تشخیص انجمن وزن‌دار می‌شود.

۵- ۳- کارهای آینده

مسئلهٔ پیش‌بینی پیوند مسئله‌ای است که همچنان نیاز به پیشرفت دارد و تلاش‌های بیشتری را برای ارائهٔ روش‌های دقیق‌تر با کارایی بالاتر می‌طلبد. نتایج پژوهش انجام گرفته نشان داد که از اطلاعات وزن‌پیوندها و انجمن‌ها، می‌توان بهره جست تا به کارایی بهتری در این روش‌ها برسیم. به عنوان کارهایی که می‌توان در آینده به آن‌ها توجه داشت، موارد زیر را می‌توان مطرح کرد:

- در راستای روش‌های ارائه شده، می‌توان تلاش کرد تا با مکانیزم‌های دیگری از اطلاعات انجمن‌های شبکه استفاده کرد. برای مثال ممکن است بتوان فرمول متفاوت و پیچیده‌تری برای توابع SM' و CO که در فصل سه معرفی شدند استفاده کرد که بتوان به کارایی بالاتری رسید.

- موارد دیگری که می‌توانند به پیش‌بینی پیوند کمک کند، استفاده از اطلاعات منحصر به فرد یال‌ها و گره‌ها به همراه هم، استفاده از دیگر ویژگی‌های ساختاری شبکه، استفاده از اطلاعات متنی در شبکه‌های اجتماعی و ترکیب این موارد با هم است.

- یکی از معضلات اصلی پیش‌بینی پیوند در شبکه‌های اجتماعی حجم بالای شبکه‌هاست که می‌تواند در مواردی، با تحمیل هزینهٔ بالا، بسیاری از روش‌های پیش‌بینی پیوند را غیرقابل اجرا کند. در بخش ۳- ۲- ۳ این پژوهش اشاره‌ای به موضوع کاهش پیچیدگی زمانی اجرای روش‌های پیشنهادی شد، اما این موضوع همچنان توجه ویژه‌ای را طلب می‌کند.

فهرست مراجع

[1] Michael PH Stumpf, Thomas Thorne, Eric de Silva, Ronald Stewart, Hyeong Jun An, Michael Lappe, and Carsten Wiuf. Estimating the size of the human interactome. *Proceedings of the National Academy of Sciences*, 105(19):6959–6964, 2008.

[2] Luis A Nunes Amaral. A truer measure of our ignorance. *Proceedings of the National Academy of Sciences*, 105(19):6795–6796, 2008.

[3] Jennifer Watling Neal. "kracking" the missing data problem: applying krackhardt's cognitive social structures to school-based social networks. *Sociology of Education*, 81(2):140–162, 2008.

[4] Christian Von Mering, Roland Krause, Berend Snel, Michael Cornell, Stephen G Oliver, Stanley Fields, and Peer Bork. Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403, 2002.

[5] Roger Guimerà and Marta Sales-Pardo. Missing and spurious interactions and the reconstruction of complex networks. *Proceedings of the National Academy of Sciences*, 106(52):22073–22078, 2009.

[6] Shi Zhou and Raúl J Mondragón. Accurately modeling the internet topology. *Physical Review E*, 70(6):066108, 2004.

[7] Shai Carmi, Shlomo Havlin, Scott Kirkpatrick, Yuval Shavitt, and Eran Shir. A model of internet topology using k-shell decomposition. *Proceedings of the National Academy of Sciences*, 104(27):11150–11154, 2007.

[8] Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. the x-random case. *International statistical review/revue internationale de Statistique*, pages 291–319, 1992.

[9] Ron Kohavi et al. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Ijcai*, volume 14, pages 1137–1145, 1995.

[10] Linyuan Lü and Tao Zhou. Link prediction in complex networks: A survey. *Physica A: Statistical Mechanics and its Applications*, 390(6):1150–1170, 2011.

[11] Peng Wang, BaoWen Xu, YuRong Wu, and XiaoYu Zhou. Link prediction in social networks: the state-of-the-art. *Science China Information Sciences*, 58(1):1–38, 2015.

[12] Prantik Bhattacharyya, Ankush Garg, and Shyhtsun Felix Wu. Analysis of user keyword similarity in online social networks. *Social network analysis and mining*, 1(3):143–158, 2011.

[13] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Effects of user similarity in social media. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 703–712. ACM, 2012.

[14] Mark EJ Newman. Clustering and preferential attachment in growing networks. *Physical Review E*, 64(2):025102, 2001.

[15] Gueorgi Kossinets. Effects of missing data in social networks. *Social networks*, 28(3):247–268, 2006.

[16] Tsuyoshi Murata and Sakiko Moriyasu. Link prediction of social networks based on weighted proximity measures. In *Web Intelligence, IEEE/WIC/ACM international conference on*, pages 85–88. IEEE, 2007.

[17] Gerard Salton and Michael J McGill. Introduction to modern information retrieval. 1986.

[18] Thorvald Sørensen. {A method of establishing groups of equal amplitude in plant sociology based on similarity of species and its application to analyses of the vegetation on Danish commons}. *Biol. Skr.*, 5:1–34, 1948.

[19] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[20] Tao Zhou, Linyuan Lü, and Yi-Cheng Zhang. Predicting missing links via local information. *The European Physical Journal B*, 71(4):623–630, 2009.

[21] EA Leicht, Petter Holme, and Mark EJ Newman. Vertex similarity in networks. *Physical Review E*, 73(2):026120, 2006.

[22] Yu-Xiao Zhu, Linyuan Lü, Qian-Ming Zhang, and Tao Zhou. Uncovering missing links with cold ends. *Physica A: Statistical Mechanics and its Applications*, 391(22):5769–5778, 2012.

[23] Lada A Adamic and Eytan Adar. Friends and neighbors on the web. *Social networks*, 25(3):211–230, 2003.

[24] Linyuan Lü and Tao Zhou. Link prediction in weighted networks: The role of weak ties. *EPL (Europhysics Letters)*, 89(1):18001, 2010.

[25] Purnamrita Sarkar, Deepayan Chakrabarti, and Andrew W Moore. Theoretical justification of popular link prediction heuristics. In *IJCAI Proceedings-International Joint Conference on Artificial Intelligence*, volume 22, page 2722, 2011.

[26] Linyuan Lü, Ci-Hang Jin, and Tao Zhou. Similarity index based on local paths for link prediction of complex networks. *Physical Review E*, 80(4):046122, 2009.

[27] Leo Katz. A new status index derived from sociometric analysis. *Psychometrika*, 18(1):39–43, 1953.

[28] Alexis Papadimitriou, Panagiotis Symeonidis, and Yannis Manolopoulos. Fast and accurate link prediction in social networking systems. *Journal of Systems and Software*, 85(9):2119–2132, 2012.

[29] Miller McPherson, Lynn Smith-Lovin, and James M Cook. Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415–444, 2001.

[30] Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543. ACM, 2002.

[31] David Liben-Nowell and Jon Kleinberg. The link-prediction problem for social networks. *Journal of the American society for information science and technology*, 58(7):1019–1031, 2007.

[32] Ryan N Lichtenwalter, Jake T Lussier, and Nitesh V Chawla. New perspectives and methods in link prediction. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 243–252. ACM, 2010.

[33] Nir Friedman, Lise Getoor, Daphne Koller, and Avi Pfeffer. Learning probabilistic relational models. In *IJCAI*, volume 99, pages 1300–1309, 1999.

[34] Han Hee Song, Tae Won Cho, Vacha Dave, Yin Zhang, and Lili Qiu. Scalable proximity estimation and link prediction in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, pages 322–335. ACM, 2009.

[35] Jorge Valverde-Rebaza and Alneu de Andrade Lopes. Exploiting behaviors of communities of twitter users for link prediction. *Social Network Analysis and Mining*, 3(4):1063–1074, 2013.

[36] Aditya Krishna Menon and Charles Elkan. Link prediction via matrix factorization. In *Machine Learning and Knowledge Discovery in Databases*, pages 437–452. Springer, 2011.

[37] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

[38] Santo Fortunato. Community detection in graphs. *Physics Reports*, 486(3):75–174, 2010.

[39] Bo Yang, Dayou Liu, and Jiming Liu. Discovering communities from social networks: methodologies and applications. In *Handbook of Social Network Technologies and Applications*, pages 331–346. Springer, 2010.

[40] Mark EJ Newman and Michelle Girvan. Finding and evaluating community structure in networks. *Physical review E*, 69(2):026113, 2004.

[41] Martin Rosvall and Carl T Bergstrom. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences*, 105(4):1118–1123, 2008.

[42] Peter D Grünwald, In Jae Myung, and Mark A Pitt. *Advances in minimum description length: Theory and applications*. MIT press, 2005.

[43] Martin Rosvall, Daniel Axelsson, and Carl T Bergstrom. The map equation. *The European Physical Journal Special Topics*, 178(1):13–23, 2010.

[44] Sucheta Soundarajan and John Hopcroft. Using community information to improve the precision of link prediction methods. In *Proceedings of the 21st international conference companion on World Wide Web*, pages 607–608. ACM, 2012.

[45] Andrea Lancichinetti, Santo Fortunato, and Filippo Radicchi. Benchmark graphs for testing community detection algorithms. *Physical review E*, 78(4):046110, 2008.

[46] Andrea Lancichinetti and Santo Fortunato. Benchmarks for testing community detection algorithms on directed and weighted graphs with overlapping communities. *Physical Review E*, 80(1):016118, 2009.

[47] Andrea Lancichinetti. LFR network benchmarks. `https://sites.google.com/site/andrealancichinetti/files`, 2009. [Online; accessed 1-August-2015].

[48] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982.

[49] Andrea Lancichinetti and Santo Fortunato. Community detection algorithms: a comparative analysis. *Physical review E*, 80(5):056117, 2009.

واژه‌نامه فارسی به انگلیسی

الف

استنتاج آماری ..... Statistical Inference
الگوریتم نقشه‌اطلاعات ... Infomap Algorithm
انجمن ..... Community
اندازهٔ برش ..... Cut Size

ب

بازیابی اطلاعات ..... Information Retrival
بالا به پایین ..... Top-down
برپایهٔ رشته ..... String-based
بر پایهٔ متن ..... Text-based
بیشینه همانندی ..... Maximum Likelihood

پ

پایین به بالا ..... Bottom-up
پیچیدگی زمانی ..... Time Complexity
پیش‌بینی پیوند ..... Link Prediction
پیوندکاوی ..... Link Mining
پیوندهای بالقوه ..... Potential Links
پیوندهای چندگانه ..... Multiple Links
پیوندهای دیده‌شده ..... Observed Links
پیوندهای گم‌شده ..... Missing Links
پیوندهای ممکن ..... Possible Links
پیوندهای موجود ..... Existent Links
پیوندهای ناموجود ..... Non-existent Links

ت

تبرید شبیه‌سازی‌شده ... Simulated Annealing
تخصیص منابع ..... Resource Allocation
تشخیص انجمن ..... Community Detection
توازن ..... Trade-off

چ

چارچوب ..... Framework

ح

حداقل طول توصیف .. Minimim Description Length

خ

خوشه‌بندی سلسله‌مراتبی ... Heirarchical Clustering

د

دقت ..... Precision

دقت n-بهترین ..... Top-n precision

دقت در n ..... Precision at n

دقت میانگین ..... Average Precision

ر

رابطهٔ دودویی ..... Binary Relation

رایانامه ..... Email

رتبه‌صفحه ..... PageRank

رتبه‌صفحه ریشه‌دار ..... Rooted PageRank

روش‌های بر پایهٔ شباهت .. Similarity-based Methods

روش‌های بر پایهٔ همبندی .. Topology-based Methods

ز

زنجیره‌های مارکوف ..... Markov Chains

س

سابقه انتشار ..... Publication Record

سامانه‌های پیشنهادگر . Recommender Systems

سوگیری ..... Bias

ش

شباهت ساختاری ..... Structural Similarity

شبکه پیچیده ..... Complex Networks

شبکهٔ مصنوعی ..... Synthesis Network

شبکهٔ همکاری بین نویسندگان .... Co-authorship Network

شبکه‌های اجتماعی برخط .. Online Social Networks

شبکه‌های دوبخشی ..... Bipartite Network

ض

ضریب تنظیم ..... Adjustment Factor

ک

کدگذاری هافمن ..... Huffman Coding

م

ماتریس مجاورت ..... Adjacency Matrix

محک ..... Bbenchmark

مرتبه ..... Order

مرکزیت ..... Centerality

مسیر محلی ..... Local Path

معیارهای بر پایهٔ گره ..... Node-based Metrics

ه

همبندی ..... Topology

همسایگان مشترک ..... Common Neighbors

و

وابستگی ترجیحی .... Prefrential Attachment

وابسته به پارامتر ..... Parameter Dependent

وبگاه ..... Website

ولگشت ..... Random Walk

واژه‌نامه انگلیسی به فارسی

A

Adjacency Matrix ..... ماتریس مجاورت

Adjustment Factor ..... ضریب تنظیم

Average Precision ..... دقت میانگین

B

Bbenchmark ..... محک

Bias ..... سوگیری

Binary Relation ..... رابطهٔ دودویی

Bipartite Network ..... شبکه‌های دوبخشی

Bottom-up ..... پایین به بالا

C

Centerality ..... مرکزیت

Co-authorship Network ..... شبکهٔ همکاری بین نویسندگان

Common Neighbors ..... همسایگان مشترک

Community ..... انجمن

Community Detection ..... تشخیص انجمن

Complex Networks ..... شبکه پیچیده

Cut Size ..... اندازهٔ برش

E

Email ..... رایانامه

Existent Links ..... پیوندهای موجود

F

Framework ..... چارچوب

H

Heirarchical Clustering .... خوشه‌بندی سلسله‌مراتبی

Huffman Coding ..... کدگذاری هافمن

I

Infomap Algorithm ..... الگوریتم نقشه‌اطلاعات

Information Retrival ..... بازیابی اطلاعات

L

Link Mining ..... پیوندکاوی

Link Prediction ..... پیش‌بینی پیوند

Local Path ..... مسیر محلی

M

زنجیره‌های مارکوف ..... Markov Chains

بیشینه همانندی ..... Maximum Likelihood

حداقل طول توصیف .. Minimim Description Length

پیوندهای گم‌شده ..... Missing Links

پیوندهای چندگانه ..... Multiple Links

N

معیارهای بر پایهٔ گره ..... Node-based Metrics

پیوندهای ناموجود ..... Non-existent Links

O

پیوندهای دیده‌شده ..... Observed Links

شبکه‌های اجتماعی برخط ... Online Social Networks

مرتبه ..... Order

P

رتبه‌صفحه ..... PageRank

وابسته به پارامتر ..... Parameter Dependent

پیوندهای ممکن ..... Possible Links

پیوندهای بالقوه ..... Potential Links

دقت ..... Precision

دقت در n ..... Precision at n

وابستگی ترجیحی ..... Prefrential Attachment

سابقه انتشار ..... Publication Record

R

ولگشت ..... Random Walk

سامانه‌های پیشنهادگر .. Recommender Systems

تخصیص منابع ..... Resource Allocation

رتبه‌صفحه ریشه‌دار ..... Rooted PageRank

S

روش‌های بر پایهٔ شباهت ... Similarity-based Methods

تبرید شبیه‌سازی‌شده ..... Simulated Annealing

استنتاج آماری ..... Statistical Inference

برپایهٔ رشته ..... String-based

شباهت ساختاری ..... Structural Similarity

شبکهٔ مصنوعی ..... Synthesis Network

T

برپایهٔ متن ..... Text-based

پیچیدگی زمانی ..... Time Complexity

بالا به پایین ..... Top-down

دقت n-بهترین ..... Top-n precision

همبندی ..... Topology

روش‌های بر پایهٔ همبندی .. Topology-based Methods

توازن ..... Trade-off

W

وبگاه ..... Website

Abstract

Nowadays, analysing social networks has became an important issue and it has attracted attentions from various fields of science. One of the most important problems here, is Link Prediction. This problem tries to predict the links that are either non-existent or unobserved. There are different approaches and methods toward this problem. Similarity-based methods is a category among them which is very popular due to its simplicity and resonable performance. Morever, in most of the previous works on this problem, link weights are not taken into account, even though they can carry valuable information. Similarly, one can use other structral information of a graph such as community information, to increase the performance of link prediction.

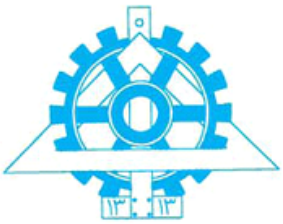
This study aims to propose a method based on community detection for link prediction in weighted networks. Briefly, the proposed method predict links inside communitie. The main reason for doing so is that its more likely for a node to establish a connection to a member of its own community, and also potetnial links inside of a community are much fewer than potential links outside of communities. This method consists of two steps and either involving or not involving the link weights in each of these steps, provide four different methods. For evaluating the performance of the proposed methods, a set of synthesis networks called LFR networks will be used which are kind of scale-free networks. After performing experiments on parameter space of these networks, we will analyze performance of the proposed methods and discuss that each of these methods can improve the performance of link prediction under what circumestances.

Keywords: Social Network Analysis, Weighted Link Prediction, Community Detection, LFR Networks

University of Tehran

College of Engineering

School of Elec. & Computer Eng.



Link Prediction in Weighted Networks

A thesis submitted to the Graduate Studies Office

in partial fulfillment of the requirements

for the degree of MS in Computer Engineering

By:

Hamid Azimy

Superviser:

Dr. Masoud Asadpour

August 2015