



ارائه یک مدل پیش‌بینی یال مبتنی بر شباهت ساختاری و هوموفیلی در شبکه‌های اجتماعی

علیرضا اسحاق‌پور^۱، مصطفی صالحی^{۲*} و وحید رنجبر^۳

^۱ گروه بین‌رشته‌ای فناوری، دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران

^۲ دانشکده مهندسی کامپیوتر، دانشگاه یزد، یزد، ایران

چکیده

در سال‌های اخیر شبکه‌های اجتماعی مجازی روز به روز در حال رشد و تغییر هستند. یال‌های جدید نشان‌دهنده تعاملات میان گره‌ها هستند و پیش‌بینی آن‌ها از اهمیت بالایی برخوردار است. معیارهای پیش‌بینی یال را می‌توان به دو گروه مبتنی بر همسایگی گره و مبتنی بر پیمایش مسیر تقسیم کرد. پژوهش‌گوان ایجاد یال جدید در شبکه را از منظر نظری به دو علم نزدیکی در گراف و هوموفیلی نسبت می‌دهند. با وجود مطالعات بسیار در حوزه علوم شبکه مطالعه تأثیر دو رویکرد نظری در کنار یک‌پیگر در ایجاد یال‌های مسئله‌ای باز محسوب می‌شود و تاکنون معیارهای شباهت مبتنی بر همسایگی گره از این منظر مطالعه نشده‌اند. در این پژوهش مدلی ارائه کردیم تا با استفاده از آن از مزایای هر دو رویکرد نزدیکی در گراف و هوموفیلی استفاده کنیم و با استفاده از آن توانستیم بر دقت معیارهای شباهت مبتنی بر همسایگی گره بیفزاییم. برای ارزیابی این پژوهش از دو مجموعه‌داده شبکه اجتماعی زنجان و شبکه اجتماعی مجازی پوک استفاده شده که مجموعه‌داده نخست برای این پژوهش جمع‌آوری و سپس تکمیل شده است.

واژگان کلیدی: پیش‌بینی یال، شباهت هوموفیلی، شباهت ساختاری، شبکه‌های اجتماعی

Providing a Link Prediction Model based on Structural and Homophily Similarity in Social Networks

Alireza Eshaghpor¹, Mostafa Salehi^{2*} & Vahid Ranjbar³

^{1,2} Faculty of New Sciences and Technologies, Tehran University, Tehran

³ Department of Computer Engineering, Yazd University, Yazd

Abstract

In recent years, with the growing number of online social networks, these networks have become one of the best markets for advertising and commerce, so studying these networks is very important. Most online social networks are growing and changing with new communications (new edges). Forecasting new edges in online social networks can give us a better understanding of the growth of these networks. Link prediction has many important applications. These include predicting future social networking interactions, the ability to manage and design useful organizational communications, and predicting and preventing relationships in terrorist gangs.

There have been many studies of link prediction in the field of engineering and humanities. Scientists attribute the existence of a new relationship between two individuals for two reasons: 1) Proximity to the graph (structure) 2) Similar properties of the two individuals (Homophilic law). Based on the two approaches mentioned, many studies have been carried out and the researchers have presented different similarity metrics for each category. However, studying the impact of the two approaches working together to create new edges remains an open problem.

* Corresponding author

*نویسنده عهده‌دار مکاتبات

Similarity metrics can also be divided into two categories; Neighborhood-based and path-based. Neighborhood-based metrics have the advantage that they do not need to access the whole graph to compute, whereas the whole graph must be available at the same time to calculate path-based metrics. So far, above the two theoretical approaches (proximity and homophily) have not been found together in the neighborhood-based metrics. In this paper, we first attempt to provide a solution to determine importance of the proximity to the graph and similar features in the connectivity of the graphs. Then obtained weights are assigned to both proximity and homophily. Then the best similarity metric in each approach are obtained. Finally, the selected metric of homophily similarity and structural similarity are combined with the obtained weights.

The results of this study were evaluated on two datasets; Zanjan University Graduate School of Social Sciences and Pokee online Social Network. The first data set was collected for this study and then the questionnaires and data collection methods were filled out. Since this dataset is one of the few Iranian datasets that has been compiled with its users' specifications, it can be of great value. In this paper, we have been able to increase the accuracy of Neighborhood-based similarity metric by using two proximity in graph and homophily approaches.

Keywords: Link prediction, Homophily similarity, Network similarity, Social networks

پیش‌بینی یال، کاربردهای بسیار مهمی دارد. از جمله آن می‌توان به پیش‌بینی تعاملات آینده شبکه‌های اجتماعی، توانایی مدیریت و طراحی ارتباطات سازمانی سودمند و پیش‌بینی روابط در باندهای ترویجی و پیش‌گیری از آن اشاره کرد [10,11]. از کاربردهای دیگر پیش‌بینی یال می‌توان به استفاده از آن برای ایجاد سامانه‌های توصیه‌گر برای دوست‌یابی در شبکه‌های اجتماعی اشاره کرد. سامانه‌های توصیه‌گر، سامانه‌هایی هستند که به کمک اطلاعات موجود و تحلیل رفتار و خصوصیات کاربران، پیشنهادهای خودکاری به آنها ارائه می‌دهند [13].

مطالعات بسیاری در رابطه با پیش‌بینی یال در حوزه مهندسی و علوم انسانی مطرح شده است، داشتن دانشمندان وجود ارتباط جدید بین دو فرد را به دو دلیل نسبت می‌دهند:

۱- نزدیکی در گراف^۱ (ساختار)،

۲- ویژگی‌های مشابه دو فرد (قانون هوموفیلی)^۲ [8-9]. بر اساس دو رویکرد بالا مطالعات بسیاری انجام شده است و پژوهش‌گران معیارهای شباهت متفاوتی را برای هر دسته ارائه کرده‌اند [21-23]. با این حال مطالعه تأثیر دو رویکرد بالا در کنار یکدیگر در ایجاد یال‌های جدید همچنان مسئله‌ای باز محسوب می‌شود.

معیارهای شباهت را می‌توان به دو گروه مبتنی بر همسایگی گره و مبتنی بر پیمایش مسیر تقسیم کرد. معیارهای مبتنی بر همسایگی گره این مزیت را دارند که برای محاسبه نیازی به دسترسی به کل گراف ندارند؛ در صورتی که برای محاسبه معیارهای مبتنی بر پیمایش مسیر، حتماً باید کل گراف دوستی در همان لحظه^(۱) در دسترس باشد [14]. تاکنون در دسته معیارهای مبتنی بر همسایگی گره دو رویکرد نظری بالا (نزدیکی در گراف و هوموفیلی) در کنار

¹ Close in network

² Homophily

۱- مقدمه

از دهه دوم قرن ۲۱، شاهدیم که مجموعه‌ای از افراد از طریق کاربردهای بستر اینترنت با یکدیگر ارتباطاتی از قبیل دوستی، ارتباط اجتماعی و همکاری برقرار می‌کنند و کنش‌های اجتماعی خود را با الهام از امکانات فنی شبکه و زمینه‌های اجتماعی و سازوکارهای ارتباطاتی سامان می‌بخشند. این مجموعه از افراد و کنش‌هایی که بین آن‌ها رخ می‌دهد، شبکه‌های اجتماعی مجازی را تشکیل می‌دهند. هرچند آمار دقیقی برای تعداد کاربران شبکه‌های اجتماعی مجازی وجود ندارد، ولی با توجه به آمار تجاری، شبکه‌های اجتماعی مجازی در سال‌های اخیر توансه‌اند تعداد کاربران بسیار بالایی را به خود جذب کنند و همین امر باعث شده است این شبکه‌ها به یکی از بهترین بازارها برای تبلیغات و تجارت تبدیل شوند [1-2]. بنابراین مطالعه این شبکه‌ها از اهمیت بالایی دارد [4-5]. پیش‌تر این شبکه‌ها، ساختار و ویژگی‌های مشخصی دارند [6]. یکی از موضوعات مورد مطالعه در علوم شبکه طراحی مدل‌هایی است که ظهور چنین ساختارهایی را پیش‌بینی کند [3, 7, 8].

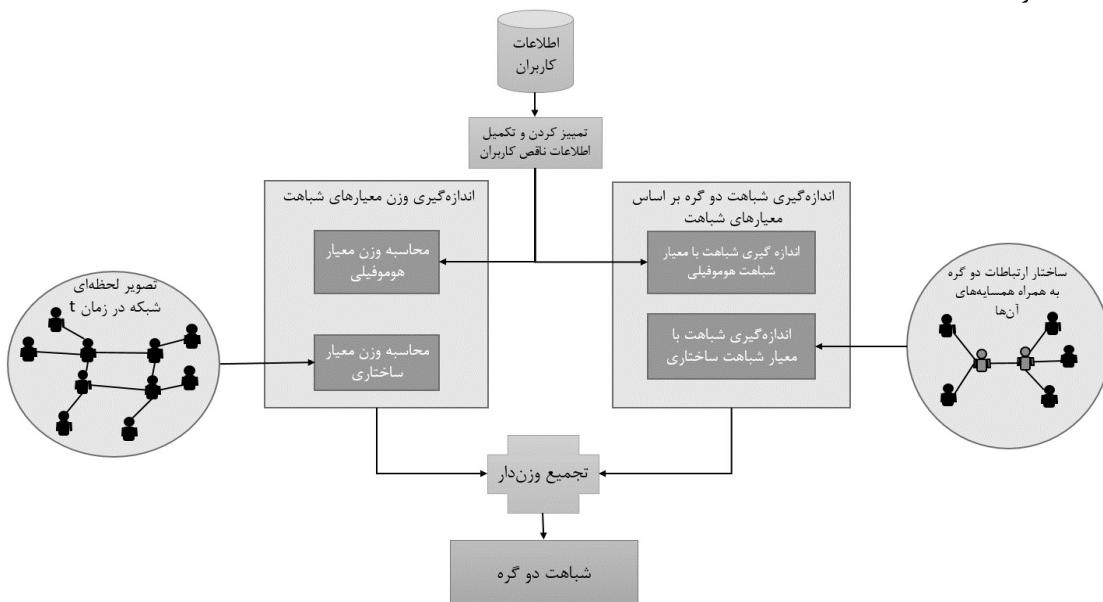
بسیاری از شبکه‌های اجتماعی مجازی، به شدت پویا بوده و با اضافه‌شدن یال‌های جدید در حال رشد و تغییرپرند. یال‌های جدید نشان‌دهنده تعاملات میان گره‌های شبکه هستند؛ بنابراین مطالعه شبکه‌های اجتماعی در سطح یال‌ها و گره‌ها می‌تواند درک بهتری از مکانیزم رشد شبکه‌ها در اختیار ما قرار دهد که به واسطه آن می‌توان به طراحی مدلی برای پیش‌بینی رشد شبکه پرداخت. این مطلب در علوم شبکه با اصطلاح پیش‌بینی یال مطرح می‌شود و به این معناست که با داشتن ساختار شبکه در لحظه t بتوان ساختار شبکه را در آینده نزدیک در زمان $x = t + t'$ پیش‌بینی کرد.



ارائه یک مدل پیش‌بینی یال مبنی بر شباهت ساختاری و هموفیلی در شبکه‌های اجتماعی

نتایج این پژوهش بر روی دو مجموعه داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان و شبکه اجتماعی مجازی پوکک ارزیابی شده است. مجموعه داده نخست برای این پژوهش جمع‌آوری شده و سپس با استفاده از پرسشنامه و روش‌های تکمیل اطلاعات، اطلاعات کاربری افراد تکمیل شده است. از آن جایی که این مجموعه داده جزو محدود مجموعه داده‌های ایرانی است که همراه با مشخصات کاربران آن جمع‌آوری شده است، می‌تواند ارزش بالایی داشته باشد.

یکدیگر دیده نشده است. در این پژوهش با ارائه یک مدل که ساختار آن در شکل (۱) ملاحظه می‌شود، توانستیم با استفاده از دو رویکرد نزدیکی در گراف و هموفیلی بر دقت معیارهای شباهت مبتنی بر همسایگی گره بیفزاییم. همچنین در مدل ارائه شده نیازی به دسترسی کل گراف در زمان ا وجود نخواهد داشت و تنها داشتن یک تصویر لحظه‌ای از شبکه در زمان گذشته (t) کفایت می‌کند؛ همچنین از خروجی این پژوهش و وزن‌هایی که برای هموفیلی شبکه و ساختار شبکه حاصل شده است، می‌توان در سایر کارهای علمی نظری انتخاب معیار تأثیرگذار هموفیلی در شبکه و تشخیص اجتماعات برخط استفاده کرد.



(شکل-۱): ساختار روش پیشنهادی
(Figure-1): The structure of the proposed approach

داده است که هر چقدر دو فرد در گراف دوستی به یکدیگر نزدیکتر باشند، در آینده نزدیک احتمال برقراری ارتباط بین آن‌ها بیش‌تر خواهد بود.

ویژگی‌های مشابه دو فرد به این نکته اشاره دارد که هر چقدر دو فرد دارای ویژگی‌های مشابه بیشتری باشند، احتمال برقراری ارتباط آن دو در آینده نزدیک بیشتر خواهد بود. از این موضوع در ادبیات علوم انسانی با عنوان قانون هموفیلی و در ادبیات علوم شبکه با عنوان اختلاط همگن^۱ یاد می‌شود. گاهی اوقات بسته به جامعه آماری، افراد تمایل دارند تا با کسانی که در یک ویژگی مشخص با آن‌ها شباهت ندارند، معاشرت کنند. برای مثال در روابط عائقبانه، افراد با کسانی

^۱ Assortative mixing

۳- پیشینه موضع

در حوزه مهندسی و علوم انسانی درخصوص پیش‌بینی یال، مطالعات زیادی انجام شده است، پژوهش گران در این مطالعات وجود ارتباط جدید بین دو فرد را به دو دلیل نسبت می‌دهند: ۱- نزدیکی در گراف - ۲- ویژگی‌های مشابه دو فرد [۹-۸]. نزدیکی در گراف به این امر اشاره دارد که دو فردی که به طور مستقیم در ارتباط نیستند، ولی به واسطه یک یا چند فرد (دوستان مشترک) در شبکه به صورت غیر مستقیم با یکدیگر ارتباط دارند، در آینده نزدیک با احتمال بیش‌تری با یکدیگر ارتباط برقرار می‌کنند (برای مثال آشنایی آن‌ها با یکدیگر در مهمنانی از طریق دوست مشترک). مطالعات نشان

از رویکردی دیگر می‌توان معیارهای شباهت را به دو دسته تقسیم کرد. ۱- معیارهای شباهت مبتنی بر همسایگی گره ۲- معیارهای شباهت مبتنی بر پیمایش مسیر. معیارهای معرفی شده در جداول (۱ و ۲) همگی در گروه معیارهای شباهت مبتنی بر همسایگی گره قرار داده می‌شوند. در دسته معیارهای مبتنی بر پیمایش مسیر که در سال‌های اخیر پژوهش‌گران به آن توجه بیشتری داشته‌اند، پیش‌بینی یال بر اساس پیمودن یک مسیر (یا فرامسیر) در گراف مشخص می‌شود. بدین معنا که برای پیش‌بینی ارتباطات آینده یک گره، از آن گره الگوریتم آغاز شده و درنهایت با پیمودن مسیری در گراف به گره‌هایی از شبکه ختم می‌شود و در پایان از بین آن‌ها تعدادی گره بر اساس اولویت به عنوان نامزد ارتباطات آینده آن گره معرفی می‌شوند [۲۶-۲۸]. بعضی از معیارهای این دسته از ویژگی‌های شباهت در پیمودن مسیر بهره برده‌اند که می‌توان گفت در این معیارها از دو رویکرد نظری بالا برای پیش‌بینی یال استفاده شده است [۸].

(جدول-۲): معیارهای شباهت هوموفیلی
(Table-2): Homophily similarity

مرجع	سال	معیار شباهت
[۱۵]	۱۹۸۶	همپوشانی
[۱۶]	۱۹۶۶	گودال
[۱۷]	۲۰۰۲	اسکین
[۱۱]	۲۰۰۸	IOF
[۱۲]	۲۰۱۳	OF

با توجه به دقت بالاتر معیارهای دسته مبتنی بر پیمایش مسیر در قیاس با معیارهای دسته مبتنی بر همسایگی گره همچنان از معیارهای مبتنی بر همسایگی گره در بسیاری از کارهای علمی استفاده می‌شود؛ علت آن است که برای محاسبه معیارهای شباهت مبتنی بر همسایگی گره، برخلاف معیارهای مبتنی بر پیمایش مسیر دسترسی به کل گراف دوستی نیاز نیست و پیچیدگی زمانی کمتری دارد؛ همین امر باعث شده است که همچنان از این معیارها استفاده شود. با توجه به مزایای معیارهای شباهت مبتنی بر همسایگی گره، رویکرد این پژوهش افزایش دقت این معیارها و تمرکز اصلی بر روی این دسته از معیارها است.

نکته قابل تأملی که وجود دارد، این است که تاکنون در بین معیارهای مبتنی بر همسایگی گره، مطالعه‌ای مبنی بر درنظرگرفتن هر دو تحیل نظری نزدیکی در گراف و هوموفیلی در کنار یکدیگر دیده نشده است. با وجود این امر در این

که جنسیت متفاوتی با آن‌ها دارند معاشرت می‌کنند، یا در جامعه خانوادگی، افراد با پدربرگ‌ها و مادربرگ‌ها که از نظر سنی با آن‌ها تفاوت زیادی دارند، ارتباط بیشتری برقرار می‌کنند. به این تمایل در ادبیات علوم شبکه اختلاط ناهماهنگ^۱ و در ادبیات علوم اجتماعی هتروفیلی^۲ گفته می‌شود.

بدیهی است، نزدیکی در گراف و هوموفیلی، به‌الزام از یکدیگر مستقل نبوده و دارای همپوشانی نیز هستند (به طور مثال بر اساس قانون هوموفیلی دوستان یک فرد با آن فرد ویژگی‌های مشترکی دارند، بنابراین وجود دوست مشترک که مبین نزدیکی در گراف است، با قانون هوموفیلی نیز، قابل توجیه است).

در حوزه مهندسی، بیش‌تر بر روی نزدیکی در گراف تمرکز شده و معیارهای متعددی بر این اساس ارائه شده که تعدادی از آن‌ها به صورت خلاصه در جدول (۱) ارائه شده است. (در ادامه معیارهای شباهتی که بر اساس نزدیکی در گراف محاسبه می‌شوند با عنوان معیارهای شباهت ساختاری عنوان می‌شوند). در مطالعات صورت‌گرفته از بین معیارهای عنوان شده، معیار جاکارد و شباهت شبکه دقت بالاتری نسبت به سایر معیارها کسب کرده‌اند [۲۵].

(جدول-۱): معیارهای شباهت ساختاری
(Table-1): Network similarity

معیار شباهت	سال	مرجع
ضریب جاکارد	۱۹۰۱	[۱۸]
همبستگی نقطه به نقطه اطلاعات مشترک	۱۹۹۱	[۱۹]
آدامیک و آدار	۲۰۰۳	[۲۰]
کسینوس	۲۰۰۴	[۲۴]
شباهت شبکه	۲۰۱۳	[۱۲]

در حوزه علوم انسانی پژوهش‌گران به مباحث نظری قوانین هوموفیلی پرداخته‌اند و با ارائه شواهد و مطالعات تجربی توانسته‌اند این مبحث را غنی‌تر کنند. با این وجود مطالعات انجام‌شده درخصوص ارائه معیارها بر اساس قوانین هوموفیلی نسبت به معیارهای شباهت ساختاری بسیار کمتر بوده است. تعدادی از معیارهای ارائه شده بر اساس قوانین هوموفیلی در جدول (۲) ارائه شده است (در ادامه به این دسته از معیارها، معیارهای شباهت هوموفیلی اطلاق می‌شود). از بین معیارهای شباهت هوموفیلی معیار OF و IOF توانسته‌اند، نتایج خوبی را کسب کنند [۱۲]. بنابراین در این پژوهش این دو معیار به عنوان معیارهای نامزد انتخاب شده‌اند.

^۱ Disassortative mixing

^۲ Hetrophily



$W_{mac}(c)$ بیشینه وزن محتمل برای ویژگی c را نشان می‌دهد و $W(c)$ بیان‌گر وزن محتمل برای ویژگی c است. $W(c)$ از تفضیل تعداد روابطی که بر اساس ویژگی هوموفیلی c در گراف دوستی حاصل شده‌اند، با تعداد روابطی که بر اساس این ویژگی هوموفیلی در گراف تصادفی و با توزیع درجه یکسان حاصل شده‌اند، به‌دست می‌آید. در این رابطه A_i ماتریس مجاورت گراف مورد مطالعه، k_i درجه گره i و m نشان‌دهنده تعداد یال‌های گراف مورد مطالعه هستند. همچنین (r_i, c_i) یک متغیر بولی است که خروجی آن در صورت شباهت ویژگی هوموفیلی دو گره i و j برابر عدد یک و در غیر این صورت برابر عدد صفر است. $W(c)$ می‌تواند مثبت، منفی و صفر باشد. می‌توانیم از بازه‌های عددی وزن حاصل، نتایج زیر را بگیریم:

$W(c) > 0$: تعداد ارتباطاتی که در این شبکه براساس شباه وجود دارد از تعداد ارتباطاتی که در حالت تصادفی به‌دست می‌آید، بیشتر است. درواقع در شبکه مورد مطالعه تعداد بیشتری از ارتباطات براساس شباه ویژگی c نسبت به مقدار مورد انتظار داریم؛ بنابراین برای ویژگی c دو نتیجه می‌توانیم بگیریم: نخست آن که در شبکه مورد مطالعه در رابطه با ویژگی c هوموفیلی وجود دارد نه هتروفیلی. برای مثال در شبکه دانشجویان، افراد با تحصیلات یکسان با یکدیگر بیشتر ارتباط برقرار می‌کنند؛ همچنین در شبکه دوستی، افرادی با گروه‌های سنی مشابه با یکدیگر ارتباط بیشتری برقرار می‌کنند. درواقع این امر بدان معناست که در این شبکه‌ها برای ویژگی تحصیلات و سن، هوموفیلی وجود دارد. دوم آن که بزرگ‌بودن مقدار عددی وزن ویژگی c نشان‌دهنده تأثیرگذاری بیشتر ویژگی c در انتخاب دوستان است.

$W(c) < 0$: تعداد ارتباطاتی که در این شبکه براساس شباه وجود دارد، از تعداد ارتباطاتی که در حالت تصادفی به‌دست می‌آید، کمتر است. درواقع در شبکه مورد مطالعه تعداد کمتری از ارتباطات براساس شباه ویژگی c نسبت به مقدار مورد انتظار داریم؛ بنابراین می‌توانیم دو نتیجه بگیریم: نخست آن که در شبکه مورد مطالعه در رابطه با ویژگی c هتروفیلی وجود دارد نه هوموفیلی. دوم آن که بزرگ‌بودن مقدار عددی وزن ویژگی c نشان‌دهنده تأثیرگذاری بیشتر ویژگی c در انتخاب دوستان است.

$W(c) = 0$: از صفر بودن وزن هوموفیلی برای ویژگی c می‌توان فهمید که این ویژگی تأثیری در انتخاب دوستان توسط افراد در شبکه مورد مطالعه ندارد؛ برای مثال (جنسیت) W برای شکل (۲) برابر $۱۲/۰$ می‌باشد، این بدین معناست که در این

پژوهش ما به‌دلیل آن هستیم تا با ارائه یک روش، علاوه‌بر حفظ مزیت روش‌های مبتنی بر همسایگی گره، بتوانیم هوموفیلی و نزدیکی در گراف را در کنار یکدیگر در نظر بگیریم تا بتوانیم بر دقت این معیارها بیافزاییم.

۳- روش اندازه‌گیری

ساده‌ترین روشی که برای درنظرگرفتن هر دو رویکرد هوموفیلی و نزدیکی در گراف وجود دارد، تجمعی معیارهای متناظر هر دو گروه است. اما نکته‌ای که در اینجا مطرح می‌شود، این است که با درنظرگرفتن اهمیت هر یک از این معیارها در شبکه مورد مطالعه، امکان گرفتن نتایج ضعیف‌تر وجود خواهد داشت.

در این پژوهش فرض شده است که ساختار شبکه در آینده نزدیک شامل تغییرات عمده‌ای نخواهد شد؛ با این فرض راه حل عملی استفاده شده در این پژوهش این است که در گام نخست ابتدا به ارزیابی شبکه مورد مطالعه پرداخته شده تا براساس آن بتوان میزان اهمیت هر یک از رویکردهای هوموفیلی و نزدیکی در گراف سنجیده شود. درواقع به‌دلیل آن هستیم که کاربران شبکه مورد مطالعه، تاکنون در انتخاب دوستان خود چقدر به مسأله نزدیکی در گراف و چقدر به مسأله وجود ویژگی‌های مشابه اهمیت داده‌اند. در ادامه براساس آن وزن‌هایی حاصل می‌شود. در نهایت معیارهای شباهت هوموفیلی و معیارهای شباهت ساختاری با یکدیگر با احتساب وزن‌هایی به دست آمده تجمعی می‌شوند. در ادامه نحوه محاسبه وزن‌ها برای هر یک از دو رویکرد هوموفیلی و نزدیکی در گراف تبیین خواهد شد.

۱-۳- اندازه‌گیری وزن ویژگی‌های هوموفیلی

نیومن، معیاری برای اندازه‌گیری ویژگی‌های هوموفیلی برای شبکه‌های اجتماعی مجازی، ارائه کرده است [29]. از این معیار می‌توان برای مشخص کردن وزن ویژگی‌های هوموفیلی برای متغیرهای اسمی استفاده شود (در صورت وجود متغیرهای ترتیبی، فاصله‌ای و نسبی، این متغیرها باید به متغیرهای اسمی تبدیل شوند)؛ بنابراین در این کار برای تعیین وزن ویژگی هوموفیلی از رابطه (۱) بهره گرفته شده است. این رابطه بیان‌گر وزن نرمال شده برای ویژگی هوموفیلی c است.

$$\frac{W(c)}{W_{max}(c)} = \frac{\sum_{ij} (A_{ij} - \frac{k_i k_j}{2m}) \delta(c_i, c_j)}{2m - \sum_{ij} \frac{k_i k_j}{2m} \delta(c_i, c_j)} \quad (1)$$



شبکه مشابهت جنسیت باعث افزایش احتمال دوستی بین افراد می‌شود (هوموفیلی) و وزن این احتمال برابر ۰/۱۲ است.

۳-۲- اندازه‌گیری وزن ویژگی ساختاری

همان‌طور که در بخش پیشینه موضع بیان شد، الگوی دوست‌یابی افراد را می‌توان به دو قانون هوموفیلی و نزدیکی در گراف، نسبت داد. در ادبیات موضع از قانون نزدیکی در گراف در بسیاری از معیارها همچون جاکارده، کسینوسی، شباهت شبکه، نرمال L1 و غیره استفاده شده است. در همه آن‌ها یک فرض ساده‌سازی به طور مستقیم استفاده شده است، و آن هم این است که در این معیارها فقط دوست مشترک به‌طور مستقیم لحاظ می‌شود و فاصله‌های دورتر مانند دوست دوست مشترک، برای دو فرد لحاظ نمی‌شوند. علت این هست که در واقعیت پژوهش‌ها نشان داده که تأثیر دوست مشترک در جذب دو فرد به یکدیگر بسیار بالاتر از تأثیر دوست‌های مشترک، با واسطه‌های بیشتر است.

فرض کنید در لحظه t در یک شبکه دو فردی که با یکدیگر ارتباط ندارند، یک یا چند دوست مشترک دارند. در لحظه $t+1$ در همان شبکه این دو فرد با یکدیگر دوست می‌شوند. دوست‌شدن دو فرد می‌تواند هم به شباهت آن دو و هم به نزدیکی آن دو در گراف دوستی وابسته باشد؛ بنابراین این دو مؤلفه مستقل از هم نیستند؛ اما با فرض ساده‌سازی استقلال آن دو از هم می‌توان ایجاد یال جدید را به نزدیکی در گراف و تأثیر دوست مشترک آن‌ها نسبت داد. این استدلال مشابه استدلال نیومن در به‌دست‌آوردن میزان تأثیر وزن هوموفیلی در شبکه که ایجاد یال بین افرادی با ویژگی‌های مشابه را مستقل از تأثیر نزدیکی در گراف دوستی آن‌ها در نظر گرفته‌اند، است [29].

با استدلال بالا می‌توان استنتاج کرد که به‌ازای هر سه‌تایی بسته در شبکه، یک یال جدید در شبکه تحت تأثیر دوست مشترک ایجاد شده است. بنابراین مشابه کار نیومن [29] برای به‌دست‌آوردن میزان اهمیت دوست مشترک و نزدیکی در گراف، کافی است تعداد سه‌تایی‌های بسته در شبکه را در قیاس با بیشینه تعداد حالات ممکن در شبکه، شمارش شود. برای این امر مطالعاتی صورت گرفت و مشخص شد برای این مسئله در ادبیات موضوع سه راه کار به‌شرح زیر وجود دارد: میانگین ضریب خوش‌بندی محلی^۱؛ ضریب خوش‌بندی محلی^۲ برای هر گره تعریف شده و کیفیت اتصالات گره یادشده با همسایه‌هایش در قیاس با یک گره در

گراف کامل^۳ را تعیین می‌کند و از رابطه (۲) برای گره مشخص نا حاصل می‌شود:

$$C_i = \frac{2e_i}{k_i(k_i-1)} \quad (2)$$

که در آن e_i برابر تعداد اتصالاتی است که بین همسایه‌های گره یادشده آ وجود دارد، همچنین k_i برابر درجه گره آ است.

فرض کنید e_i در شبکه‌ای برابر عدد k_i باشد، بنابراین k_i یال در شبکه ایجاد شده است که دو سر تمامی آن یال‌ها را همسایه‌های گره آ تشکیل داده‌اند؛ درنتیجه تمامی گره‌های دو سر K یال، دارای حداقل یک دوست مشترک هستند، که آن هم گره آ است. بنابراین براساس فرضیات بالا تمام k_i یال به‌خاطر نزدیکی در گراف و به‌خاطر واسطه‌گری گره آ در شبکه ایجاد شده‌اند؛ لذا کافی است برای محاسبه وزن ویژگی ساختاری برای هر گره این مقدار را تقسیم بر مقدار بیشینه آن کنیم، که به‌طور دقیق برابر C_i می‌شود. حال برای به‌دست‌آوردن وزن ساختاری کل شبکه کافیست میانگین C_i ‌ها گرفته شود که برابر با میانگین ضریب خوش‌بندی محلی می‌شود. رابطه (۳) نحوه محاسبه ضریب خوش‌بندی محلی را نشان می‌دهد:

$$\bar{C} = \frac{1}{n} \sum_i^n \frac{2e_i}{k_i(k_i-1)} \quad (3)$$

ضریب خوش‌بندی سراسری^۴: معیاری است که میزان اتصالات کل گره‌های گراف را نسبت به گراف کامل متناظر آن می‌دهد. ضریب خوش‌بندی سراسری از رابطه (۴) حاصل می‌شود:

$$C = \frac{\text{تعداد مثلث‌ها در گراف} \times ۳}{\text{تعداد اتصالات سه‌تایی در گراف}} \quad (4)$$

موتیف سه‌تایی بسته^۵: در علوم شبکه مطالعاتی در رابطه با تعداد همبندی‌های m_{tai} در گراف با عنوان موتیف انجام می‌شود که در آن تعداد همبندی m_{tai} در شبکه با تعداد همبندی همان شبکه در حالت تصادفی و با حفظ توزیع درجه، قیاس می‌شود. درواقع موتیف به‌دبیال پیداکردن فرکانس الگوهای ساختاری شبکه نسبت به حالت تصادفی است. فرمول محاسبه موتیف، در رابطه (۵) آورده شده است. موتیف را به صورت $Motif_M(e)$ نشان می‌دهند، که در آن M تعداد گره‌های همبندی مورد مطالعه و e تعداد یال‌های همبندی موردنظر موتیف است. $n_M(e)$ تعداد M تایی‌هایی که e یال دارند، در گراف اصلی می‌شمارد؛ همچنین گراف تصادفی متناظر با گراف اصلی با حفظ درجه به صورت تصادفی چندین

³ Clique (Complete graph)

⁴ Global Clustering Coefficient

⁵ Motif (3)

شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان: این شبکه اجتماعی مجازی، کار خود را از سال ۸۹ شروع کرده و تا سال ۹۳ در فضای مجازی در در دسترس بوده است. حدود ششصد دانشجو در دانشگاه تحصیلات تکمیلی زنجان مشغول به تحصیل بوده‌اند که از این بین، حدود سیصد نفر در شبکه اجتماعی مجازی این دانشگاه، به عضویت این تارنما در آمدند. تمامی اطلاعات این وبسایت (در طول سال‌های ۸۹ تا ۹۲) جمع‌آوری شده است که شامل ۳۰۱ کاربر (شامل دانشجویان، استادی و کارمندان) به همراه مشخصات واردشده و همچنین ۳۱۸۲ ارتباطات بین آن‌ها است.

تکمیل اطلاعات مشخصات کاربران

بسیاری از مجموعه‌داده‌های جمع‌آوری شده از شبکه‌های اجتماعی مجازی به‌دلایل مختلف از قبیل پر نکردن تمام مقادیر پرسش‌نامه توسط کاربران در هنگام ثبت‌نام و یا فعل سازی سطح حریم خصوصی توسط کاربران (که در این صورت APIها قادر به استخراج اطلاعات این‌گونه افراد نیستند) دارای نقصان اطلاعات هستند. همین موضوع باعث شده است تا تکمیل اطلاعات ناقص در شبکه‌های اجتماعی مجازی اهمیت پیدا کند. از آنجا که در مجموعه‌داده جمع‌آوری شده نقصان مشخصات کاربران وجود داشت، بنابراین ابتدا از طریق پرسش‌نامه مشخصات مربوطه کامل‌تر شد. در ادامه برای تکمیل اطلاعات مشخصات کاربران از یکی از روش‌های تکمیل اطلاعات ناقص کاربران در شبکه‌های اجتماعی در [12] بهره گرفته شد. در این روش با استفاده از قانون هموفیلی اطلاعات ناقص تکمیل می‌شوند. روش کار به این صورت هست که اطلاعات ناقص ویژگی آنم (برای مثال جنسیت) در حساب کاربری فرد x ، به واسطه مقادیر ویژگی آنم در حساب کاربری دوستان مشترک آن فرد، تکمیل می‌شود. این کار با رأی اکثریت انجام می‌شود. درواقع هر دوست فرد x ، یک رأی دارد؛ درنهایت برای رفع نقص ویژگی آنم در حساب کاربری فرد x صفتی که بیشترین رأی را کسب کرده است، انتخاب می‌شود. این درحالی است که دو عامل محدود‌کننده برای پذیرش صفت وجود دارد:

۱. کمترین تعداد رای‌ها (f): برای جلوگیری از نتیجه گیری به‌ازای تعداد آرای خیلی کم از یک عامل محدود‌کننده برای تعداد آرای استفاده شده است که با f نمایش داده شده است. درصورتی که تعداد رأی‌های صفت برگزیده از یک حدی کمتر باشد، صفت پذیرش نمی‌شود (ویژگی ۱ برای گره x همچنان خالی باقی می‌ماند).

^۱ <http://coinlab.ut.ac.ir/resources>

مرتبه ساخته می‌شود و میانگین و انحراف معیار همبندی‌های M تابی که e یال دارند، در آن شمرده می‌شود که در رابطه (۵) به‌ترتیب با $n_M(e)^{Random} < n_M(e)$ نشان داده شده است [30]:

$$Motif_M(e) = \frac{n_M(e) - \langle n_M(e)^{Random} \rangle}{\sigma_M^{Random}(e)} \quad (5)$$

از آنجا که میانگین ضریب خوشبندی محلی به‌ازای هر دوست مشترک، تعداد یال‌های ایجادشده تحت تأثیر آن را نسبت به بیشینه حالت ممکن شمارش می‌کند، لذا نزدیک‌ترین روش از سه روش بالا، این روش است؛ درنهایت فرمول وزن ویژگی ساختاری به‌صورت رابطه (۳) محاسبه می‌شود:

بنابراین با توجه به شکل (۱) محاسبه وزن معیار هموفیلی با دراختیارداشتن اطلاعات کاربران توسط رابطه (۱) و محاسبه وزن معیار ساختاری با دراختیارداشتن تصویر لحظه‌ای شبکه در زمان t توسط رابطه (۳) به‌دست می‌آید. بنابراین فرمول نهایی تجمعی وزن دار در رابطه (۶) در زیر ارائه شده است.

$$P(X, Y) = \begin{cases} \frac{\frac{W(c)}{W_{max}(c)} + (W_{structural} \times NS(u, x))}{\frac{W(c)}{W_{max}(c)} + W_{structural}}, & if X_k = Y_k \\ \frac{\left(\frac{W(c)}{W_{max}(c)} \times S_k(X_k, Y_k) \right) + (W_{structural} \times NS(u, x))}{\frac{W(c)}{W_{max}(c)} + W_{structural}}, & Otherwise \end{cases} \quad (6)$$

در این رابطه $W_{structural}$ برابر رابطه (۴)، $NS(u, x)$ برابر معیار شباهت شبکه و $S_k(X_k, Y_k)$ برابر معیار شباهت Ω است. X_k نیز مقدار ویژگی هموفیلی X برای گره X و همچنین $P(X, Y)$ برابر احتمال ایجاد یال بین دو گره X و Y است.

۴- نتایج

در این بخش ابتدا در رابطه با نحوه جمع‌آوری و تکمیل دو مجموعه داده مورد استفاده در این پژوهش، توضیحاتی داده شده است، در ادامه به تبیین سناریوی ارزیابی و درنهایت به تحلیل و ارزیابی نتایج حاصل از دو مجموعه داده پرداخته شده است.

۴-۱- مجموعه داده

برای این پژوهش دو مجموعه داده در نظر گرفته شده است که در ادامه هر یک تبیین خواهد شد.

شبکه اجتماعی دانشگاه تحصیلات تكمیلی زنجان در جدول (۳) قابل مشاهده است. ویژگی جنسیت به علت نداشتن نقص و کامل بودن مقادیر آن برای تمام افراد در این جدول وارد نشده است. دقت هر ویژگی به ازای بهترین مقادیری که برای کرو / محاسبه شده به دست آمده است، همان طور که در این جدول ملاحظه می شود، به علت دقت پایین تکمیل اطلاعات ویژگی شهرستان محل سکونت این ویژگی در نظر گرفته نشده و بر روی آن اجرا گرفته نشده است. ستون «تعداد مقادیر ناقص قبل از اجرا» در این جدول، نشان دهنده تعداد نقص هر ویژگی پس از تکمیل اطلاعات به واسطه پرسش نامه است. ستون چهارم در این جدول، نشان دهنده دقت روش اجرایی برای تکمیل اطلاعات ناقص است که همان طور که مشخص است، دقت قابل قبولی به ازای تمام ویژگی ها به استثنای ویژگی شهرستان محل سکونت به ازای این روش حاصل شده است. از آنجا که در این پژوهش به دنبال بالابردن دقت روش نمونه برداری پیوند خودکار هستیم، و اطلاعات ناقص بر روی نتایج نهایی تأثیرگذار است، تنها دو ویژگی «جنسیت» و «وضعیت تأهل» که درنهایت به صورت کامل و بدون اطلاعات ناقص در شبکه حاصل شده اند، برای این پژوهش در نظر گرفته شده است.

۲. بیشینه قاطع در رأی گیری (۱): برای جلوگیری از نتیجه گیری به ازای درصد های پایین شرکت کننده ها، یک عامل محدود کننده برای درصد آرا استفاده شده است که با / مشخص شده است. درصورتی که درصد آرا به ازای تعداد افراد شرکت کننده در رأی گیری از یک حدی کمتر باشد آن صفت مورد قبول واقع نمی شود.

اگر به f و t مقادیری کمی داشته باشند، اطلاعات بیشتری تخمین زده می شود؛ ولی دقت روش کاهش می یابد، همچنین اگر مقدار f و t زیاد باشد، اطلاعات کمی تخمین زده می شود. بنابراین انتخاب مقادیر f و t حائز اهمیت هستند. براساس رابطه (۷) بهینه ترین حالت برای این مقادیر در مجموعه داده آموزشی به دست می آید.

$$L(f,t) = Precision(f,t) \times Log(C(f,t)) \quad (7)$$

که در آن $Precision(f,t)$ دقت تخمین به واسطه مقادیر f و t در مجموعه داده آموزشی و همین طور $C(f,t)$ تعداد تخمین های درست در مجموعه داده آموزشی است. با مقدار دهی مختلف تابع بالا و ارزیابی بهترین نتیجه به صورت تجربی مقادیر بهینه برای f و t محاسبه می شوند. مقادیر f و t برای بیشینه L بهینه ترین مقادیر ممکن هستند. نتایج حاصل از شبیه سازی این روش تکمیل اطلاعات برای مجموعه داده

(جدول-۳): نتایج تکمیل اطلاعات ناقص مجموعه داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان
(Table-3): Result of complete missing data on IASBS Social

	f	t	تعداد مقادیر پیش بینی شده	دقت	تعداد مقادیر ناقص قبل از اجرا	درصد مقادیر ناقص باقی مانده بعد از اجرا
وضعیت تأهل	2	0.1	166	0.95	167	0
قطع تحصیلی	1	0.5	81	0.79	131	16.5
سال ورود به دانشگاه	3	0.1	94	0.64	164	23
رشته تحصیلی	1	0.5	89	0.89	156	22
شهرستان محل سکونت	1	0.3	این ویژگی به خاطر دقت پایین در نظر گرفته نشد.	0.45	125	-

به علت حجم بودن این مجموعه داده یک نمونه برداری جستجوی نخست سطح با شروع از قدیمی ترین گره از این مجموعه داده انجام شد و ۴۷۲۴۱ گره از این مجموعه داده به همراه ۸۹۴۷۷۶ یال از آن در این نمونه برداری، برداشت شد (برداشت با قدیمی ترین گره آغاز شده است، بدین ترتیب می توان فرض کرد که شبکه حاصل از نمونه برداری، شبکه اجتماعی مجازی پوکک در سال های اولیه است).

شبکه اجتماعی مجازی پوکک^۱: پوکک محبوب ترین شبکه اجتماعی مجازی در کشور اسلواکی است که حتی پس از آمدن فیسبوک، محبوبیت این تارنما نزول پیدا نکرده است. مجموعه داده این شبکه اجتماعی مجازی در طول ده سال جمع آوری شده است که شامل یک میلیون و شصصد هزار کاربر می شود [31]. این مجموعه داده بر روی تارنماه دانشگاه استنفورد^۲ قرار داده شده است.

¹ Pokec

² این مجموعه داده در پیوند: <https://snap.stanford.edu/data/soc-pokec.html> در دسترس است.



بیشتر از احتمال به دست آمده برای یال ایجاد نشده بود، آن گاه:

$$n' = n' + 1$$

v. در غیر این صورت اگر احتمال‌های به دست آمده با هم برابر بودند، آن گاه:

$$n'' = n'' + 1$$

v. با داشتن n و n' مقدار AUC محاسبه می‌شود.

$$AUC = \frac{n' + 0.5n''}{n}$$

vi. مقدار مجموع AUC‌ها برای اجراهای متفاوت محاسبه می‌شود:

$$AUC_{Total} = AUC_{Total} + AUC$$

2. از ده اجرای گرفته شده میانگین گرفته و مقدار نهایی AUC حاصل می‌شود:

$$AUC_{Final} = \frac{AUC_{Total}}{10}$$

برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان در قسمت b از الگوریتم بالا به علت داشتن زمان تشکیل یال‌ها به جای حذف تصادفی یال‌ها، ده درصد از آخرين یال‌های تشکیل شده حذف شده‌اند؛ همچنین برای هر دو مجموعه‌داده ۱۱ برابر نصف تعداد یال‌های حذف شده در نظر گرفته شده است (در واقع ۱۱ برابر است با پنج درصد از کل یال‌های گراف اصلی).

۳-۴- نتایج و تحلیل داده

در ابتدا برای انتخاب معیار شباهت اصلاح از میان معیارهای شباهت نامزد، مطالعه‌ای انجام دادیم که نتایج آن در زیر آورده شده است.

۱-۴- انتخاب معیار اصلاح از میان معیارهای نامزد شباهت

نتایج بر روی مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان برای انتخاب معیار شباهت هموفیلی برگزیده از بین دو معیار نامزد OF و IOF در جدول (۴) و برای انتخاب معیار شباهت ساختاری برگزیده از بین چهار معیار نامزد شباهت شبکه، کسینوس، نرمال L1 و جاکارد در جدول (۵) آورده شده است. همچنین نتایج مشابه بر روی مجموعه‌داده شبکه اجتماعی مجازی پوکک به ترتیب در جداول (۶) و (۷) آورده شده است.

برای به دست آوردن نتایج از معیار ارزیابی AUC که در بخش قبل توضیح داده شد، استفاده شده است. با این تفاوت که به جای استفاده از معیار شباهت تجمیع شده در آن از

این مجموعه‌داده دارای ویژگی‌های زیادی است ولی مقادیر بیشتر آن‌ها بیش از پنجاه درصد دارای نقصان هستند و تنها ویژگی که تمامی مقادیر برای آن وجود دارد، ویژگی جنسیت است. بنابراین از آنجا که در این پژوهش اطلاعات ناقص بر روی نتایج نهایی تأثیرگذار است، تنها همین ویژگی از این مجموعه‌داده در نظر گرفته شده است.

۲-۴- سناریوی ارزیابی

اصلی‌ترین معیارهای ارزیابی استفاده شده در حوزه پیش‌بینی یال عبارتند از: دقت^۱، فراخوانی^۲، معیار F^۳، صحت رده‌بندی^۴، مساحت زیرنمودار^۵.

معیار صحت رده‌بندی به مسئله این پژوهش مربوط نمی‌شود، معیارهای دقت و فراخوانی و به طبع آن معیار F به علت این که برای محاسبه، همه یال‌های مشاهده‌نشده را بررسی می‌کنند، از نظر پیچیدگی زمانی، پیچیدگی بسیار بالایی دارند (هر سه معیار پیاده‌سازی شدن، ولی در عمل به علت زمان بر بودن، از استفاده از این معیارها صرف نظر شد) که برای مجموعه‌داده‌های بزرگ با چالش همراه خواهد شد. بنابراین از بین معیارهای بالا، معیار مساحت زیرنمودار برای ارزیابی روش پیشنهادی انتخاب شد. از این معیار در ارزیابی بسیاری از کارهای علمی استفاده شده است [14, 32, 33]

سناریوی ارزیابی به این شرح است:

- a. الگوریتم زیر به تعداد ده مرتبه انجام می‌شود.
- b. تمام یال‌هایی که در گراف ایجاد نشده‌اند، در گروه یال‌های ایجاد نشده قرار داده می‌شوند.
- c. ده درصد از یال‌های شبکه به صورت تصادفی حذف و در گروه یال‌های مشاهده‌نشده قرار داده می‌شوند.
- d. وزن ویژگی‌های ساختاری و هموفیلی برای نود درصد باقی‌مانده گراف، محاسبه می‌شوند.
- e. به تعداد ۱۱ بار روال زیر اجرا می‌شود:

 - f. به صورت تصادفی و مستقل یک یال از مجموعه یال‌های ایجاد نشده و یک یال از مجموعه یال‌های مشاهده‌نشده انتخاب می‌شود.
 - ii. براساس وزن‌های حاصل شده، احتمال یال منتخب از مجموعه یال‌های ایجاد نشده و مشاهده‌نشده براساس تجمعی وزن دار معیارهای شباهت محاسبه می‌شوند.
 - iii. اگر احتمال به دست آمده برای یال مشاهده نشده

¹ Precision

² Recall

³ FMeasure

⁴ Classification Accuracy

⁵ Area Under Curve (AUC)



معیارهای یادشده استفاده شده است. همان‌طور که مطرح شد به علت تصادفی بودن الگوریتم به تعداد ده اجرا بهازی هر یک از معیارهای شباهت معیار AUC محاسبه شده، و درنهایت از آن میانگین گرفته شده است.

(جدول-۴): ارزیابی معیار اصلاح از بین معیارهای نامزد شباهت هموفیلی برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی علوم پایه زنجان.

(Table-4): The results of the evaluation of best similarity metrics between homophily similarity metrics on IASBS Social

میانگین AUC های به دست آمده	معیار شباهت ساختاری	صفت مورد مطالعه	میانگین AUC های به دست آمده
OF	شباهت شبکه	جنسيت	0.55
	جاكارد	وضعیت تأهل	0.52
IOF	شباهت کوسمیتوسی	جنسيت	0.53
	L1	وضعیت تأهل	0.52

۴-۳-۲- ارزیابی و تحلیل نتایج

در این بخش توسط معیار ارزیابی مطرح شده، به ارزیابی و تحلیل نتایج می‌پردازیم. ویژگی‌های شباهت به دو دسته هموفیلی و ساختاری تقسیم می‌شوند. دسته ویژگی ساختاری تنها شامل یک ویژگی است؛ اما در دسته ویژگی‌های هموفیلی، تعداد ویژگی‌ها برابر با تعداد مشخصاتی است که از کاربران شبکه اجتماعی در دسترس است. تخصیص وزن به معیارهای شباهت و تجمیع آن‌ها را از دو جهت می‌توان

مطالعه کرد که در زیر به آن اشاره می‌شود:

- زمانی که چندین ویژگی کاربران یک شبکه اجتماعی در دسترس باشد، برای استفاده از معیار شباهت هموفیلی باید نتایج هر یک محاسبه و با یکدیگر تجمیع شوند. چالشی که در اینجا مطرح می‌شود، این است که آیا اهمیت هر یک از ویژگی‌ها در انتخاب دوست تأثیر یکسانی دارد؟ در بسیاری از مطالعات صورت گرفته وزن هر یک از ویژگی‌های هموفیلی را یکسان در نظر گرفته‌اند. این امر در صورتی است که مطالعات اخیر نشان داده است، این امر مبتنی بر واقعیت نیست [34-36]؛ بنابراین می‌توان به واسطه وزن‌های به دست آمده برای ویژگی‌های هموفیلی به تجمیع وزن دار آن‌ها برداخت و نتایج حاصل آن را با نتایج میانگین‌گیری ساده مقایسه کرد. نتایج حاصل به‌واسطه معیار ارزیابی مطرح شده بر روی مجموعه‌داده شبکه اجتماعی مجازی دانشگاه علوم پایه زنجان در جداول (۸) و (۹) ارائه شده است (با توجه به در دسترس بودن تنها یک ویژگی هموفیلی برای مجموعه‌داده شبکه اجتماعی مجازی پوک، از این مجموعه‌داده در این بخش استفاده نشده است).

با توجه به نتایج حاصل از ارزیابی که در جداول (۴) و (۶) ارائه شده، معیار شباهت OF توانسته است دقت بالاتری نسبت به معیار IOF کسب کند؛ بنابراین در این کار پژوهشی از بین معیارهای شباهت هموفیلی، این معیار انتخاب شده است.

(جدول-۵): ارزیابی معیار اصلاح از بین معیارهای نامزد شباهت ساختاری برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی علوم پایه زنجان.

(Table-5): The evaluation of best similarity metrics between network similarity metrics on IASBS Social

معیار شباهت ساختاری	میانگین AUC های به دست آمده
شباهت شبکه	0.56
جاكارد	0.56
شباهت کوسمیتوسی	0.53
L1	0.51

(جدول-۶): ارزیابی معیار اصلاح از بین معیارهای نامزد شباهت هموفیلی برای مجموعه‌داده شبکه اجتماعی مجازی پوک

(Table-6): The evaluation of best similarity metrics between network similarity metrics on IASBS Social

صفت مورد مطالعه	معیار شباهت	میانگین AUC های به دست آمده
OF	جنسيت	0.49
IOF	جنسيت	0.46

همان‌طور که در جداول (۵) و (۷) شان داده شده است، دو معیار شباهت شبکه و جاكارد، توانستند بیشترین دقت را نسبت به سایر معیارها در هر دو مجموعه‌داده کسب کنند

تجمیع ویژگی‌های هوموفیلی و ساختاری کمک کند، که این امر از اهمیت بیشتری برخوردار بوده و هدف اصلی این پژوهش است. راه کار پیشنهادی به این شرح است که ابتدا هر یک از ویژگی‌های هوموفیلی و ساختاری با معیار شباهت متناظر خود محاسبه و سپس به‌واسطه وزن‌های حاصل شده بهازای هر یک از آن‌ها با یکدیگر تجمیع شوند. همان‌طور که عنوان شد در ادبیات موضوع برای مسأله پیش‌بینی یال دو رویکرد وجود دارد (نژدیکی درگراف و هوموفیلی)، با استفاده از راه کار بالا می‌توان از مزایای هر دو رویکرد متناظر با درجه اهمیت هر یک بهره‌مند شد.

(جدول-۱۰): نتایج تجمیع معیارهای هوموفیلی و ساختاری برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحقیقات تکمیلی علوم پایه.

(Table-10): Evaluation results for aggregation homophily features and network features on IASBS Social.

میانگین AUC	میانگینها
0.57	تجمیع ساده (جنسیت، وضعیت تأهل و ساختاری)
0.60	تجمیع وزن دار (جنسیت، وضعیت تأهل و ساختاری)

در جدول (۱۰) نتایج تجمیع معیارها به صورت وزن دار (راه کار پیشنهادی) و میانگین‌گیری ساده برای شبکه زنجان، ارائه شده است. روش پیشنهادی توانسته است، نسبت به سایر نتایج (معیار شباهت جنسیت به‌تهابی)، معیار شباهت وضعیت تأهل به‌تهابی، تجمیع ساده دو معیار یادشده و تجمیع دو معیار یادشده به‌صورت وزن دار) دقت را 0.03 افزایش دهد. همان‌طور که عنوان شد نژدیکی وزن‌های حاصله باعث شده است که میزان افزایش دقت قابل ملاحظه نشود؛ درواقع میزان افزایش دقت مسأله وابسته به میزان تفاوت وزن ویژگی‌های حاصل شده (میزان تفاوت اهمیت ویژگی‌های شبکه) است.

همان‌طور که در جدول (۱۱) مشاهده می‌شود، وزن ویژگی هوموفیلی جنسیت برای مجموعه داده شبکه اجتماعی مجازی پوکک منفی شده است و همان‌طور که توضیح داده شد وزن منفی نشان‌دهنده وجود هتروفیلی برای این ویژگی در این شبکه است؛ درواقع در این شبکه افراد تمایل بیشتری به برقراری ارتباط با جنس مخالف از خود نشان داده‌اند. از آنجا که معیار ارزیابی پیاده‌سازی شده برای حالت تصادفی برابر است با مقدار پنجاه درصد، لذا انتظار می‌رود که (با توجه به این که در این شبکه هتروفیلی وجود دارد) برای معیار شباهت هوموفیلی جنسیت این مقدار زیر پنجاه درصد حاصل شود.

همان‌طور که در جدول (۹) مشاهده می‌شود، نتایج حاصل بر روی شبکه اجتماعی مجازی دانشگاه تحقیقات تکمیلی زنجان حاکی از آن است که با تخصیص وزن‌ها به ویژگی‌های جنسیت وضعیت تأهل و استفاده از آن در تجمیع آن‌ها برای به‌دست آوردن معیار شباهت، توانسته‌ایم بر دقت این روش نسبت به زمانی که، وزن هر معیار مشابه در نظر گرفته شده است، 0.02 بیفزاییم. از آنجا که وزن‌های حاصل شده اختلاف ناچیزی با یکدیگر دارند، درنظر گرفتن وزن‌ها در تجمیع در قیاس با حالت پایه (درنظر گرفتن وزن پکسان برای هر دو معیار) بهبود اندکی مشاهده شده است. به طور طبیعی هر چه اختلاف وزن‌ها برای معیارهای شباهت هوموفیلی بیشتر باشند، اختلاف دقت نتیجه روش ارائه شده با روش پایه بیشتر خواهد بود. همچنین از آنجا که وزن‌های تخصیص داده شده به‌ترتیب به شباهت ساختاری، شباهت هوموفیلی برای ویژگی جنسیت و شباهت هوموفیلی برای ویژگی وضعیت تأهل، بیشترین اعداد را نسبت داده است، انتظار داشتیم که معیار ارزیابی نیز بیشترین دقت را ابتدا به ویژگی ساختاری سپس به ویژگی هوموفیلی جنسیت و درنهایت به ویژگی هوموفیلی وضعیت تأهل نسبت دهد. که نتایج منطبق بر همین امر است.

(جدول-۸): نتایج ارزیابی بهازای هر ویژگی شباهت برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحقیقات تکمیلی علوم پایه.

(Table-8): Evaluation results for each similarity feature on IASBS Social

میانگین وزن‌ها	میانگین AUC
شباهت ساختاری (NS)	0.48
شباهت هوموفیلی (جنسیت)	0.29
شباهت هوموفیلی (وضعیت تأهل)	0.24

(جدول-۹): نتایج ارزیابی تجمیع ویژگی‌های هوموفیلی برای مجموعه‌داده شبکه اجتماعی مجازی دانشگاه تحقیقات تکمیلی علوم پایه.

(Table-9): Evaluation results for aggregation homophily features on IASBS Social.

میانگین AUC‌های به‌دست آمده	میانگین AUC‌های
تجمیع ساده دو معیار هوموفیلی	0.55
تجمیع وزن دار دو معیار هوموفیلی	0.57

- محاسبه وزن‌های ویژگی‌های هوموفیلی و ساختاری علاوه بر حل چالش بالا (که توضیح داده شد)، می‌تواند در

روش‌های تکمیل اطلاعات در حوزه علوم شبکه این نقص اطلاعات را برطرف کردیم.

براساس مطالعات صورت‌گرفته، معیارهای شباهتی از هر دو دسته معیارهای شباهت هوموفیلی و ساختاری به عنوان نامزد در نظر گرفتیم و با ارزیابی نتایج روی دو مجموعه داده، بهترین معیارها را برگزیدیم.

در این پژوهش توانستیم با ارائه یک مدل به تجمعی وزن دار معیارهای شباهت هوموفیلی و ساختاری بپردازیم. همچنین به واسطه آن توانستیم به دقت بهترین نتایج معیارهای مبتنی بر همسایگی گره بیافراییم. از مزایای این کار پژوهشی بی‌نیازی به دسترسی به کل گراف در همان لحظه است و فقط تصویر لحظه‌ای شبکه در زمان گذشته می‌تواند کفایت کند.

در شاخه علوم انسانی به علت تعدد زیاد ویژگی‌های هوموفیلی انتخاب تعدادی از ویژگی‌های هوموفیلی خود یکی از مسائل و چالش‌ها محسوب می‌شود. از نتایج این پژوهش می‌توان در حل این مشکل استفاده کرد. همچنین در حوزه تشخیص اجتماعات روش‌های وجود دارد که با استفاده از هوموفیلی و نزدیکی در گراف به تشخیص اجتماعات می‌پردازد. با استفاده از نتایج این پژوهش و تشخیص وزن‌های هوموفیلی و نزدیکی در گراف می‌توان به دقت این روش‌ها کمک کرد. در این پژوهش معیاری برای هتروفیلی معرفی نشد که می‌توان به عنوان کارهای آینده به آن پرداخت و نتایج این پژوهش را غنی تر کرد.

6- References

۶- مراجع

- [1] d. boyd and N. B. Ellison, "Social network sites: Definition, history, and scholarship," *Journal of Computer-Mediated Communication*, vol. 13, no. 1, pp. 210-230, 2007.
- [2] H. Gangadharbatla, "Facebook me: Collective self-esteem, need to belong, and internet self-efficacy as predictors of the iGeneration's attitudes toward social networking sites," *Journal of interactive advertising*, vol. 8, no. 2, pp. 5-15, 2008.
- [3] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, p. 167–256, 2003.
- [4] V. Ranjbar, M. Salehi, P. Jandaghi, and M. Jalili. "QANet: Tensor Decomposition Approach for Query-based Anomaly Detection in Heterogeneous Information Networks." *IEEE Transactions on Knowledge and Data Engineering*, 31(11), pp.2178-2189, 2019.

(جدول-۱۱): نتایج وزن‌های حاصل شده به ازای هر ویژگی

شباهت برای مجموعه داده پوک

(Table-11): Results of the weights obtained per similarity feature on Pokec Social

	میانگین وزن‌ها AUC
شباهت ساختاری	0.072 0.79
شباهت هوموفیلی (جنسیت)	-0.039 0.49

همان‌طور که در جدول (۱۱) مشاهده می‌شود، مقدار مساحت زیرنمودار برای ویژگی هوموفیلی جنسیت زیر پنجاه درصد حاصل شده است که مبتنی با تحلیل نظری ارائه شده است. از آنجا که معیارهای هوموفیلی جهت اندازه‌گیری هوموفیلی ارائه شده‌اند و کارایی برای هتروفیلی نداشته، اضافه کردن ویژگی هوموفیلی به ساختاری به طور طبیعی باید دقیق را کاهش دهد. همان‌طور که در جدول (۱۱) مشاهده می‌شود، میزان مساحت زیرنمودار محاسبه شده برای ویژگی ساختاری بtentheای برابر $0.79/79$ است که با اضافه شدن معیار هوموفیلی به صورت وزن دار این عدد به $0.68/68$ کاهش می‌یابد. با این حال همان‌طور که در جدول (۱۲) مشاهده می‌شود، اگر اضافه کردن معیار هوموفیلی جنسیت به صورت میانگیری ساده، اضافه شود، مقدار $0.79/79$ به $0.65/65$ کاهش می‌یابد. این امر بدان معناست که با احتساب این که در شبکه هتروفیلی وجود دارد، ولی میزان اهمیت آن از ویژگی ساختاری کمتر بوده (کمتر بودن وزن ویژگی هوموفیلی جنسیت از ویژگی ساختاری در جدول ۱۱) بنابراین در تجمعی وزن دار از میزان اهمیت آن نسبت به میانگین‌گیری ساده کاسته شده و توانسته است، نزول کمتری پیدا کند.

(جدول-۱۲): نتایج تجمعی معیارهای هوموفیلی و ساختاری برای مجموعه داده پوک

(Table-12): Evaluation results for aggregation homophily features and network features on Pokec Social.

	میانگین وزن‌ها AUC
تجمعی ساده (معیار هوموفیلی و ساختاری)	0.65
تجمعی وزن دار (معیار هوموفیلی و ساختاری)	0.68

۵- نتیجه‌گیری

در این پژوهش در ابتدا مجموعه داده شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان را جمع‌آوری کردیم. این مجموعه داده دارای نقص اطلاعات کاربران بود. برای حل آن از طریق پرسش‌نامه و درنهایت به‌واسطه پیاده‌سازی یکی از



- Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150-1170, 2011.
- [15] C. S. a. D. Waltz, "Toward memory based reasoning," ACM, vol. 29, no. 12, pp. 1213-1228, 1986.
- [16] E. Eskin, A. Arnold, M. Prerau, L. Portnoy and S. Stolfo, "A geometric framework for unsupervised anomaly detection: Detecting intrusions in unlabeled data," Applications of data mining in computer security, vol. 6, pp. 77-102, 2002.
- [17] D. W. Goodall, "A New Similarity Index Based on Probability," Biometrics, vol. 22, no. 4, pp. 882-907, 1966.
- [18] T. J. Cover TM, Elements of information theory, New York: Wiley-Interscience, 1991.
- [19] P. Jaccard, Etude de la distribution florale dans une portion des Alpes et du Jura, Naturelles: Bulletin de la Societe Vaudoise des Sciences Naturelles, 1901.
- [20] E. A. L. A. Adamic, "Friends and neighbors on the Web," Social Networks, vol. 25, no. 3, pp. 211-230, 2003.
- [21] M. Jalili, Y. Orouskhani, M. Asgari, N. Alipourfard and M. Perc, "Link prediction in multiplex online social networks," Royal Society Open Science, vol. 4, no. 2, 2017.
- [22] Z. Wu, Y. Lin and J. Wang, "Link prediction with node clustering coefficient," Physica A: Statistical Mechanics and its Applications, vol. 452, pp. 1-8, 2016.
- [23] Sh .Najari, M. Salehi, V. Ranjbar, and M. Jalili, "Link prediction in multiplex networks based on interlayer similarity." Physica A: Statistical Mechanics and its Applications, 2019..
- [24] M. Deshpande and G. Karypis, "Item-based top-n recommendation algorithms," ACM Transactions on Information Systems (TOIS), vol. 22, no. 1, pp. 143-177, 2004.
- [25] T. Zhou, L. Lü and Y.-C. Zhang, "Predicting missing links via local information," The European Physical Journal B-Condensed Matter and Complex Systems, vol. 71, no. 4, pp. 623-630, 2009.
- [26] H. Shakibian and N. Moghadam Charkari, "Mutual information model for link prediction in heterogeneous complex networks," Scientific Reports, vol. 7, 2017.
- [27] H. Zhao, Q. Yao, J. Li, Y. Song and D. Lee, "Meta-Graph Based Recommendation Fusion over Heterogeneous Information Networks," in Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2017.
- [28] X. Yang, T. Deng, Z. Guo and . Z. Ding, "Advertising Keyword Recommendation based on Supervised Link Prediction in Multi-Relational Network," in Proceedings of the [5] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," Reviews of Modern Physics, vol. 74, no. 47, January 2002.
- [6] C. Zhong, M. Salchi, S. Shah, M. Cobzarencu, N. Sastry and M. Cha, "Social bootstrapping: how pinterest and last. fm social communities benefit by borrowing links from facebook," in Proceedings of the 23rd international conference on World wide web, 2014.
- [7] S. N. Dorogovtsev and J. F. F. Mendes, Evolution of networks: From biological nets to the Internet and WWW, Oxford University Press, 2003.
- [8] J. Leskovec and L. Backstrom, "Supervised random walks: predicting and recommending links in social networks," in Proceedings of the fourth ACM international conference on Web search and data mining, 2011.
- [9] M. McPherson, L. Smith-Lovin and J. M Cook, "Birds of a Feather Homophily in Social Networks," jstor, vol. 27, pp. 415-444, 2001.
- [10] م. صالحی، ح. عبداللهان و ع. اسحاقپور، "هموفیلی در شبکه‌های اجتماعی مجازی (مطالعه موردی شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان)،" فصلنامه مطالعات رسانه‌های نوین، جلد ۶، صفحه ۹۱ تا ۹۵، ۱۳۹۵.
- [11] A. Eshaghpoor, M. Salehi and H. Abdollahyan, "Homophily in Virtual Social Networks: A Case Study of Virtual Social Network in Graduate University of Zanjan," New media studies, vol. 2, no. 6, pp. 93-120, 2016.
- [12] S. Boriah, V. Chandola and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," Red, vol. 30, no. 2, pp. 3, 2008.
- [13] C. G. Akcora, B. Carminati and F. Ferrari, "User similarities on social networks," Social Network Analysis and Mining, vol. 3, no. 3, pp. 475-495, 2013.
- [14] م. حسینی، م. ناصرالهی، ع. بقایی، "یک سامانه توصیه‌گر ترکیبی با استفاده از اعتماد و خوشبندی دوچهته به منظور افزایش کارایی بالایش گروهی"، فصل نامه پژوهش علوم و داده‌ها، جلد ۱۵، شماره ۲، صفحه ۱۱۹ تا ۱۳۲، ۱۳۹۷.
- [15] Hosseini M, Nasrollahi M, Baghaei A. "A hybrid recommender system using trust and bi-clustering in order to increase the efficiency of collaborative filtering". JSRP.; vol.15 (2) . pp.119-132, 2018.
- [16] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A:

دریافت درجه دکترا در همین رشته از دانشگاه شریف شد. پس از آن پسادکترا خود را بهترتبی در سالهای ۱۳۹۴ و ۱۳۹۵ در دانشگاه بولونیا کشور ایتالیا و دانشگاه تلکام سود پاریس (فرصت مطالعاتی) گذراند. وی هم اکنون به عنوان عضو هیئت علمی گروه بین رشته‌ای فناوری دانشگاه تهران، با مرتبه دانشیاری مشغول به فعالیت است. زمینه‌های پژوهشی ایشان شامل شبکه‌های اجتماعی مجازی، اینترنت اشیا و شبکه‌های چندلایه است.

نشانی رایانمه ایشان عبارت است از:
mostafa_salchi@ut.ac.ir



وحید رنجبر کارشناسی و کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات بهترتبی در سالهای ۱۳۹۰ و ۱۳۹۲ به پایان رسانید. وی در سال ۱۳۹۷ دکترای تخصصی خود را در رشته فناوری اطلاعات در دانشگاه تهران دریافت کرد. وی هم اکنون به عنوان عضو هیئت علمی دانشکده مهندسی کامپیوتر دانشگاه یزد مشغول به فعالیت است. زمینه‌های پژوهشی مورد علاقه ایشان تحلیل شبکه‌های اطلاعاتی، یادگیری ماشین و کلان داده است.

نشانی رایانمه ایشان عبارت است از:
vranjbar@yazd.ac.ir

26th International Conference on World Wide Web Companion, Perth, Australia, 2017.

- [29] M. E. J.33 Newman, Networks An Introduction, New York: Oxford University Press Inc, 2010.
- [30] R. Milo , S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii and U. Alon, "Network motifs: simple building blocks of complex networks," Science, vol. 298, no. 5594, pp. 824-827, 2002.
- [31] L. Takac and M. Zabovsky, "Data Analysis in Public Social Networks," in International Scientific Conference and International Workshop Present Day Trends of Innovations, 2012
- [32] K.-L. Goh, B. Kahng and D. Kim, "Universal Behavior of Load Distribution in Scale-Free Networks," Physical Review Letters, vol. 87, no. 27, 2001.
- [33] M. McPherso, L. Smith-Lovin and J. M Cook, "Classification of scale free networks," Proceedings of the National Academy of Sciences of the United States of America, vol. 99, 2002.
- [34] J. A Smith, M. McPherson and L. Smith-Lovin, "Social Distance in the United States Sex, Race, Religion, Age, and Education Homophily among Confidants, 1985 to 2004," American Sociological Review, vol. 79, no. 3, pp. 432-456, 1 6 2014.
- [35] X. Han, "Mining user similarity in online social networks: analysis, modeling and applications," 2015.
- [36] J. Hyung Kang and K. Lerman, "Using Lists to Measure Homophily on Twitter," in Workshops at the Twenty-Sixth AAAI Conference on Artificial Intelligence, 2012.



علیرضا اسحاقپور متولد ۱۳۶۹ کارشناسی و کارشناسی ارشد خود را بهترتبی در سالهای ۱۳۹۲ و ۱۳۹۶ در دانشگاه IASBS و دانشگاه تهران به پایان رسانید. وی هم اکنون به عنوان کارشناس ارشد SOC با هلدینگ بهین راه کار همکاری می کند. زمینه‌های پژوهشی مورد علاقه ایشان شبکه‌های اجتماعی مجازی، الگوریتم‌های ژنتیک چندهدفه و امنیت اطلاعات است.

نشانی رایانمه ایشان عبارت است از:
a_eshaghpoor@ut.ac.ir



مصطفی صالحی کارشناسی و کارشناسی ارشد خود را در رشته مهندسی کامپیوتر بهترتبی در سالهای ۱۳۸۴ و ۱۳۸۶ به پایان رسانید. او در سال ۱۳۹۱ موفق به