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

علیرضا اسحاق‌پور[1]، مصطفی صالحی[2*] و وحید رنجبر[3]

[1و2] گروه بین‌رشته‌ای فناوری، دانشکده علوم و فنون نوین، دانشگاه تهران، تهران، ایران

[3] 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 نشان‌دهندهٔ 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. پژوهش‌گران ایجاد یال جدید در شبکه را از منظر نظری به دو علت نزدیکی در گراف و هوموفیلی نسبت می‌دهند. با وجود مطالعات بسیار در حوزه علوم شبکه مطالعه تأثیر دو رویکرد نظری در کنار یکدیگر در ایجاد یال‌ها مسأله‌ای 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 مجموعه‌دادهٔ شبکه اجتماعی مجازی زنجان و شبکه اجتماعی مجازی پوکک استفاده شده که مجموعه‌دادهٔ 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

Alircza Eshaghpour[1], Mostafa Salchi[2*] & Vahid Ranjbar[3]

[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 (Homophile 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

*نویسندهٔ عهده‌دار مکاتبات

● تاریخ ارسال: ۹۷/۰۲/۰۳ ● تاریخ پذیرش: ۹۸/۰۴/۱۹ ● تاریخ انتشار: ۹۸/۱۲/۲۸

[DOI: 10.29252/jsdp.16.4.45]

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 homophile) 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 homophile. 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 Pokec 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

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 توانسته‌اند 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] بیش‌تر این شبکه‌ها، ساختار و ویژگی‌های مشخصی دارند [5-6]. 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، به‌شدت پویا بوده و با اضافه‌شدن یال‌های جدید در حال رشد و تغییرند. یال‌های جدید نشان‌دهندهٔ 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 آیندهٔ نزدیک در زمان $t' = t + x$ پیش‌بینی کرد.

پیش‌بینی یال، کاربردهای بسیار مهمی دارد. از جمله آن می‌توان به پیش‌بینی تعاملات آیندهٔ شبکه‌های اجتماعی، توانایی مدیریت و طراحی ارتباطات سازمانی سودمند و پیش‌بینی روابط در باندهای تروریستی 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 به‌کمک اطلاعات موجود و تحلیل رفتار و خوصوصیات کاربران، پیشنهادهای خودکاری به آنها ارائه می‌دهند [13].

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

۱- نزدیکی در گراف[1] (ساختار)

۲- ویژگی‌های مشابه دو فرد (قانون هوموفیلی[2]) [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 مسأله‌ای 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؛ درصورتی‌که برای محاسبه معیارهای مبتنی بر پیمایش مسیر، حتما باید کل گراف دوستی در همان لحظه (t) در دسترس باشد [14].

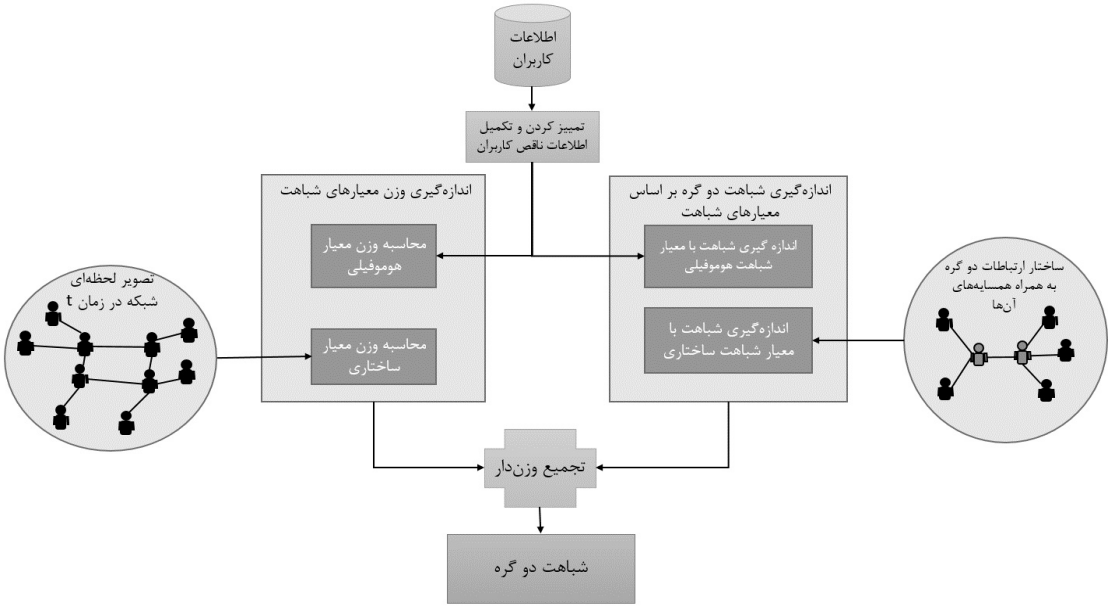
تاکنون در دسته معیارهای مبتنی بر همسایگی گره دو رویکرد نظری بالا (نزدیکی در گراف و هوموفیلی) در کنار

[1] Close in network
[2] Homophily

[DOI: 10.29252/jsdp.16.4.45]

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 هوموفیلی بر دقت معیارهای شباهت مبتنی بر همسایگی گره بیفزاییم. همچنین در مدل ارائه شده نیازی به دسترسی کل گراف در زمان t 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 هوموفیلی شبکه و ساختار شبکه حاصل شده است، می‌توان در سایر کارهای علمی نظیر انتخاب معیار تأثیرگذار هوموفیلی در شبکه و تشخصیص اجتماعات برخط استفاده کرد.

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



(شکل-۱): ساختار روش پیشنهادی

(Figure-1): The structure of the proposed approach

۲- پیشینهٔ موضوع

در حوزهٔ مهندسی و علوم انسانی درخصوص پیش‌بینی یال، مطالعات زیادی انجام شده است، پژوهش‌گران در این مطالعات وجود ارتباط جدید بین دو فرد را به دو دلیل نسبت می‌دهند: ۱- نزدیکی در گراف ۲- ویژگی‌های مشابه دو فرد [9-8]. نزدیکی در گراف به این امر اشاره دارد که دو فردی که به‌طور مستقیم در ارتباط نیستند، ولی به واسطه یک یا چند فرد (دوستان مشترک) در شبکه به‌صورت غیر مستقیم با یکدیگر ارتباط دارند، در آیندهٔ نزدیک با احتمال بیش‌تری با یکدیگر ارتباط برقرار می‌کنند (برای مثال آشنایی آن‌ها با یکدیگر در مهمانی از طریق دوست مشترک). مطالعات نشان داده است که هر چقدر دو فرد در گراف دوستی به یکدیگر نزدیک‌تر باشند، در آیندهٔ 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 هوموفیلی و در ادبیات علوم شبکه با عنوان اختلاط همگن[1] 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 عاشقانه، افراد با کسانی

[1] Assortative mixing

[DOI: 10.29252/jsdp.16.4.45]

که جنسیت متفاوتی با آن‌ها دارند معاشرت می‌کنند؛ یا در جامعه خانوادگی، افراد با پدربزرگ‌ها و مادربزرگ‌ها که از نظر سنی با آن‌ها تفاوت زیادی دارند، ارتباط بیشتری برقرار می‌کنند. به این تمایل در ادبیات علوم شبکه اختلاط ناهماهنگ[1] و در ادبیات علوم اجتماعی هتروفیلی[2] 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 حوزهٔ مهندسی، بیش‌تر بر روی نزدیکی در گراف تمرکز شده و معیارهای متعددی بر این اساس ارائه شده که تعدادی از آن‌ها به‌صورت 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 عنوان‌شده، معیار جاکارد و شباهت شبکه دقت بالاتری نسبت به سایر معیارها کسب کرده‌اند [25].

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

معیار شباهت	سال	مرجع
ضریب جاکارد	1901	[18]
همبستگی نقطه به نقطه اطلاعات مشترک	1991	[19]
آدامیک و آدار	2003	[20]
کسینوس	2004	[24]
شباهت شبکه	2013	[12]

در حوزهٔ 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 براساس قوانین هوموفیلی نسبت به معیارهای شباهت ساختاری بسیار کمتر بوده است. تعدادی از معیارهای ارائه‌شده 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 [12]. بنابراین در این پژوهش این دو معیار به‌عنوان معیارهای نامزد انتخاب شده‌اند.

[1] Disassortative mixing
[2] Hetrophily

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 معرفی‌شده 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 آیندهٔ 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 به‌عنوان نامزد ارتباطات آینده آن گره معرفی می‌شوند [28-26]. 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 [8].

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

معیار شباهت	سال	مرجع
هم‌پوشانی	1986	[15]
گودال	1966	[16]
اسکین	2002	[17]
IOF	2008	[11]
OF	2013	[12]

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؛ علّت آن است که برای محاسبه معیارهای شباهت مبتنی بر همسایگی گره، بر خلاف معیارهای مبتنی بر پیمایش مسیر دسترسی به کل گراف دوستی نیاز نیست و پیچیدگی زمانی کمتری دارند؛ 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.

نکتهٔ 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

[DOI: 10.29252/jsdp.16.4.45]

پژوهش ما به‌دنبال 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، تجمیع معیارهای متناظر هر دو گروه است. اما نکته‌ای که در اینجا مطرح می‌شود، این است که با درنظرنگرفتن اهمیت هر یک از این معیارها در شبکهٔ مورد مطالعه، امکان گرفتن نتایج ضعیف‌تر وجود خواهد داشت.

در این پژوهش فرض شده است که ساختار شبکه در آیندهٔ نزدیک شامل تغییرات عمده‌ای نخواهد شد؛ با این فرض راه حل عملی استفاده‌شده 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 به‌دنبال 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 وزن‌های به دست آمده تجمیع می‌شوند. در ادامه نحوه محاسبه وزن‌ها برای هر یک از دو رویکرد هوموفیلی و نزدیکی در گراف تبیین خواهد شد.

۳-۱- اندازه‌گیری وزن ویژگی‌های هوموفیلی

نیومن، معیاری برای اندازه‌گیری ویژگی‌های هوموفیلی برای شبکه‌های اجتماعی مجازی، ارائه کرده است [29]. از این معیار می‌توان برای مشخص‌کردن 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 نرمال‌شده برای ویژگی هوموفیلی c است.

$$\frac{W(c)}{W_{max}(c)} = \frac{\sum_{ij}(A_{ij}-\frac{k_i k_j}{2m})\delta(c_i,c_j)}{2m-\sum_{ij}\frac{k_i k_j}{2m}\delta(c_i,c_j)} \quad (۱)$$

$W_{mac}(c)$ 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 $A_{i,j}$ ماتریس مجاورت گراف مورد مطالعه، k_i درجه گره i و m نشان‌دهندهٔ تعداد یال‌های گراف مورد مطالعه هستند. همچنین $\delta(c_i, c_j)$ یک متغیر بولی است که خروجی آن در صورت شباهت ویژگی هوموفیلی دو گره i و j برابر عدد یک و در غیر این‌صورت 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 تشابه وجود دارد از تعداد ارتباطاتی که در حالت تصادفی به‌دست می‌آید، بیشتر است. درواقع در شبکه مورد مطالعه تعداد بیشتری از ارتباطات براساس تشابه ویژگی c نسبت به مقدار مورد انتظار داریم؛ بنابراین برای ویژگی c دو نتیجه می‌توانیم بگیریم: نخست آن‌که در شبکهٔ مورد مطالعه در رابطه با ویژگی c، هوموفیلی وجود دارد نه هتروفیلی. برای مثال در شبکهٔ دانشجویان، افراد با تحصیلات یکسان با یکدیگر بیشتر ارتباط برقرار می‌کنند؛ همچنین در شبکهٔ 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 آن‌که بزرگ‌بودن مقدار عددی وزن ویژگی c نشان‌دهندهٔ تأثیرگذاری بیشتر ویژگی c در انتخاب دوستان است.

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

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

[DOI: 10.29252/jsdp.16.4.45]

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

۲-۳- اندازه‌گیری وزن ویژگی ساختاری

همان‌طور که در بخش پیشینهٔ موضوع بیان شد، الگوی دوست‌یابی افراد را می‌توان به دو قانون هوموفیلی و نزدیکی در گراف، نسبت داد. در ادبیات موضوع از قانون نزدیکی در گراف در بسیاری از معیارها همچون جاکارد، 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، کافی‌ست 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 مسأله در ادبیات موضوع سه راه‌کار به‌شرح زیر وجود دارد:

میانگین ضریب خوشه‌بندی محلی[1] : ضریب خوشه‌بندی محلی[2] برای هر گره تعریف شده و کیفیت اتصالات گره یادشده با همسایه‌هایش در قیاس با یک گره در گراف کامل[3] را تعیین می‌کند و از رابطه (۲) برای گره مشخص i حاصل می‌شود:

$$C_i = \frac{2e_i}{k_i(k_i-1)} \qquad (۲)$$

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

فرض کنید e_i در شبکه‌ای برابر عدد k باشد، بنابراین k یال در شبکه ایجاد شده است که دو سر تمامی آن یال‌ها را همسایه‌های گره i تشکیل داده‌اند؛ درنتیجه تمامی گره‌های دو سر K یال، دارای حداقل یک دوست مشترک هستند، که آن هم گره i است. بنابراین براساس فرضیات بالا تمام k یال به‌خاطر نزدیکی در گراف و به‌خاطر واسطه‌گری گره i 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 به‌طوردقیق برابر C_i می‌شود. حال برای به‌دست‌آوردن وزن ساختاری کل شبکه کافی‌ست میانگین C_iها گرفته شود که برابر با میانگین ضریب خوشه‌بندی محلی می‌شود. رابطه (۳) نحوه محاسبه ضریب خوشه‌بندی محلی را نشان می‌دهد:

$$\bar{C} = \frac{1}{n}\sum_i^n \frac{2e_i}{k_i(k_i-1)} \qquad (۳)$$

ضریب خوشه‌بندی سراسری[4]: معیاری است که میزان اتصالات کل گره‌های گراف را نسبت به گراف کامل متناظر آن می‌دهد. ضریب خوشه‌بندی سراسری از رابطه (۴) حاصل می‌شود:

$$C=\frac{\text{تعداد مثلث‌ها در گراف} \times 3}{\text{تعداد اتصالات سه تایی در گراف}} \qquad (۴)$$

موتیف سه‌تایی بسته[5]: در علوم شبکهٔ مطالعاتی در رابطه با تعداد هم‌بندی‌های mتایی در گراف با عنوان موتیف انجام می‌شود که در آن تعداد همبندی mتایی در شبکه با تعداد هم‌بندی همان شبکه در حالت تصادفی و با حفظ توزیع درجه، قیاس می‌شود. درواقع موتیف به‌دنبال پیداکردن فرکانس الگوهای ساختاری شبکه نسبت به حالت تصادفی است. فرمول محاسبه موتیف، در رابطه (۵) آورده شده است. موتیف را به‌صورت $Motif_M(e)$ نشان می‌دهند، که در آن M تعداد گره‌های هم‌بندی مورد مطالعه و e تعداد یال‌های هم‌بندی مورد مطالعه 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 به‌صورت تصادفی چندین

[1] Average Local Cluster Coefficient
[2] Local Clustring Cofficient
[3] Clique (Complete graph)
[4] Global Clustering Coefficient
[5] Motif (3)

[DOI: 10.29252/jsdp.16.4.45]

مرتبه ساخته می‌شود و میانگین و انحراف معیار هم‌بندی‌های Mتایی که e یال دارند، در آن شمرده می‌شود که در رابطهٔ (۵) به‌ترتیب با $n_M(e)^{Random} >$ و $\sigma_M^{Random}(e)$ نشان داده شده است [30]:

$$Motif_M(e) = \frac{n_M(e) - <n_M(e)^{Random}>}{\sigma_M^{Random}(e)} \quad (۵)$$

از آنجا که میانگین ضریب خوشه‌بندی 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 هوموفیلی 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 تجمیع وزن‌دار در رابطه (۶) در زیر ارائه شده است.

$$P(X,Y) = \begin{cases} \frac{\frac{W(c)}{W_{max}(c)} + (W_{Structural} \times NS(u,x))}{\frac{W(c)}{W_{max}(c)} + W_{Structural}} & , if\ X_k = Y_k \\ \frac{\left(\frac{W(c)}{W_{max}(c)} \times S_k(X_k, Y_k)\right) + (W_{Structural} \times NS(u,x))}{\frac{W(c)}{W_{max}(c)} + W_{Structural}} & , Otherwise \end{cases} \quad (۶)$$

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 OF است. X_k نیز مقدار ویژگی هوموفیلی 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- مجموعه‌داده

برای این پژوهش دو مجموعه‌داده در نظر گرفته شده است که در ادامه هر یک تبیین خواهند شد.

شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان[1]: 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 ۹۳) جمع‌آوری شده است که شامل ۳۰۱ کاربر (شامل دانشجویان، اساتید و کارمندان) به‌همراه مشخصات واردشده و همچنین ۳۱۸۲ ارتباطات بین آن‌ها است.

تکمیل اطلاعات مشخصات کاربران

بسیاری از مجموعه‌داده‌های جمع‌آوری‌شده از شبکه‌های اجتماعی مجازی به‌دلایل مختلف از قبیل پر نکردن تمام مقادیر پرسش‌نامه توسط کاربران در هنگام ثبت‌نام و یا فعال‌سازی 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 جمع‌آوری‌شده 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 هوموفیلی اطلاعات ناقص تکمیل می‌شوند. روش کار به این صورت هست که اطلاعات ناقص ویژگی iام (برای مثال جنسیت) در حساب کاربری فرد x، به واسطهٔ مقادیر ویژگی iام 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 iام 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 نتیجه‌گیری به‌ازای تعداد آرای خیلی کم از یک عامل محدودکننده برای تعداد آرا استفاده شده است که با f نمایش داده شده است. درصورتی‌که تعداد رأی‌های صفت برگزیده از یک حدی کمتر باشد، صفت پذیرش نمی‌شود (ویژگی i برای گره x همچنان خالی باقی می‌ماند).

[1] http://coinlab.ut.ac.ir/resources

[DOI: 10.29252/jsdp.16.4.45]

۲. بیشینه قاطع در رأی‌گیری (t): برای جلوگیری از نتیجه‌گیری به‌ازای درصدهای پایین شرکت‌کننده‌ها، یک عامل محدودکننده برای درصد آرا استفاده شده است که با t مشخص شده است. درصورتی‌که درصد آرا به‌ازای تعداد افراد شرکت‌کننده در رأی‌گیری از یک حدی کمتر باشد آن صفت مورد قبول واقع نمی‌شود.

اگر به f و t مقادیری کمی داشته باشند، اطلاعات بیشتری تخمین زده می‌شود؛ ولی دقت روش کاهش می‌یابد، همچنین اگر مقدار fو t زیاد باشد، اطلاعات کمی تخمین زده می‌شود. بنابراین انتخاب مقادیر f و t حائز اهمیت هستند. براساس رابطه (۷) بهینه‌ترین حالت برای این مقادیر در مجموعه دادهٔ آموزشی به‌دست می‌آید.

$$L(f,t) = Precision(f,t) \times Log\big(C(f,t)\big) \qquad (۷)$$

که در آن $Precision(f,t)$ دقت تخمین به واسطه مقادیر f و t در مجموعه‌دادهٔ آموزشی و همین‌طور $C(f,t)$ تعداد تخمین‌های درست در مجموعه داده آموزشی است. با مقداردهی متفاوت تابع بالا و ارزیابی بهترین نتیجه به‌صورت تجربی مقادیر بهینه برای fو t محاسبه می‌شوند. مقادیر fو t برای بیشینه L بهینه‌ترین مقادیر ممکن هستند. نتایج حاصل از شبیه‌سازی این روش تکمیل اطلاعات برای مجموعه‌داده شبکه اجتماعی دانشگاه تحصیلات تکمیلی زنجان در جدول (۳) قابل مشاهده است. ویژگی جنسیت به‌علت نداشتن نقص و کامل‌بودن مقادیر آن برای تمام افراد در این جدول وارد نشده است. دقت هر ویژگی به‌ازای بهترین مقادیری که برای f و t محاسبه شده به‌دست آمده است؛ همان‌طورکه در این جدول ملاحظه می‌شود، به‌علت 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، نشان‌دهندهٔ تعداد نقص هر ویژگی پس از تکمیل اطلاعات به واسطه پرسش‌نامه است. ستون چهارم در این جدول، نشان‌دهنده دقت روش اجرایی برای تکمیل اطلاعات ناقص است که همان‌طور که مشخص است، دقت قابل قبولی به‌ازای تمام ویژگی‌ها به‌استثنای ویژگی شهرستان محل سکونت به‌ازای این روش حاصل شده است. از آنجا که در این پژوهش به‌دنبال بالابردن دقت روش نمونه‌برداری پیوند خودکار هستیم، و اطلاعات ناقص بر روی نتایج نهایی تأثیرگذار است، تنها دو ویژگی «**جنسیت**» و «**وضعیت تأهل**» که درنهایت به‌صورت کامل و بدون اطلاعات ناقص در شبکه حاصل شده‌اند، برای این پژوهش در نظر گرفته شده است.

(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 پیش‌بینی شده	دقت	تعداد مقادیر ناقص قبل از اجرا	درصد مقادیر ناقص باقی مانده بعد از اجرا
وضعیت تاهل	2	0.1	166	0.95	167	0
مقطع 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 درنظر گرفته نشد.	0.45	125	-

شبکهٔ اجتماعی مجازی پوکـــک[1]: پوکک محبوب‌ترین شبکه اجتماعی مجازی در کشور اسلواکی است که حتی پس از آمدن فیس‌بوک، محبوبیت این تارنما نزول پیدا نکرده است. مجموعه‌داده این شبکه اجتماعی مجازی در طول ده سال جمع‌آوری شده است که شامل یک‌میلیون و ششصد هزار کاربر می‌شود [31]. این مجموعه‌داده بر روی تارنمای دانشگاه استنفورد[2] قرار داده شده است.

به‌علت حجیم‌بودن این مجموعه‌داده یک نمونه‌برداری جستجوی نخست سطح با شروع از قدیمی‌ترین گره از این مجموعه‌داده انجام شد و ۴۷۲۴۱ گره از این مجموعه‌داده به‌همراه ۸۹۴۷۷۶ یال از آن در این نمونه‌برداری، برداشت شد (برداشت با قدیمی‌ترین گره آغاز شده است؛ بدین ترتیب می‌توان فرض کرد که شبکهٔ حاصل از نمونه‌برداری، شبکه اجتماعی مجازی پوکک در سال‌های اولیه است).

[1] Pokec

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

[DOI: 10.29252/jsdp.16.4.45]

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 استفاده‌شده در حوزهٔ پیش‌بینی یال عبارتند از: دقت[1]، فراخوانی[2]، معیار F[3]، صحت رده‌بندی[4]، مساحت زیرنمودار[5].

معیار صحت رده‌بندی به مسأله این پژوهش مربوط نمی‌شود، معیارهای دقت و فراخوانی و به‌طبع آن معیار F به‌علت این‌که برای محاسبه، همهٔ یال‌های مشاهده‌نشده را بررسی می‌کنند، از نظر پیچیدگی زمانی، پیچیدگی بسیار بالایی دارند (هر سه معیار پیاده‌سازی شدند، ولی درعمل به‌علت زمان‌بر بودن، از استفاده از این معیارها صرف نظر شد) که برای مجموعه‌داده‌های بزرگ با چالش همراه خواهند 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:

۱. الگوریتم زیر به تعداد ده مرتبه انجام می‌شود.

a. تمام یال‌هایی که در گراف ایجاد نشده‌اند، در گروه یال‌های ایجادنشده قرار داده می‌شوند.

b. ده درصد از یال‌های شبکه به‌صورت تصادفی حذف و در گروه یال‌های مشاهده‌نشده قرار داده می‌شوند.

c. وزن ویژگی‌های ساختاری و هوموفیلی برای نود درصد باقی‌ماندهٔ گراف، محاسبه می‌شوند.

d. به تعداد n بار روال زیر اجرا می‌شود:

i. به‌صورت تصادفی و مستقل یک یال از مجموعه یال‌های ایجادنشده و یک یال از مجموعه یال‌های مشاهده‌نشده انتخاب می‌شود.

ii. براساس وزن‌های حاصل‌شده، احتمال یال منتخب از مجموعه یال‌های ایجادنشده و مشاهده‌نشده براساس تجمیع وزن‌دار معیارهای شباهت محاسبه می‌شوند.

iii. اگر احتمال به‌دست‌آمده برای یال مشاهده نشده بیشتر از احتمال به‌دست‌آمده برای یال ایجادنشده بود، آن‌گاه:

$$n' = n' + 1$$

iv. در غیر این‌صورت اگر احتمال‌های به‌دست‌آمده با هم برابر بودند، آن‌گاه:

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

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

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

e. مقدار مجموع AUCها برای اجراهای متفاوت محاسبه می‌شود:

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

۲. از ده اجرای گرفته‌شده میانگین گرفته و مقدار نهایی AUC حاصل می‌شود:

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

برای مجموعه‌داده شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان در قسمت b از الگوریتم بالا به‌علت داشتن زمان تشکیل یال‌ها به‌جای حذف تصادفی یال‌ها، ده درصد از آخرین یال‌های تشکیل‌شده حذف شده‌اند؛ همچنین برای هر دو مجموعه‌داده n برابر نصف تعداد یال‌های حذف‌شده در نظر گرفته شده است (درواقع n برابر است با پنج درصد از کل یال‌های گراف اصلی).

۳-۴- نتایج و تحلیل داده

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

۱-۳-۴- انتخاب معیار اصلح از میان معیارهای نامزد شباهت

نتایج بر روی مجموعه‌داده شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان برای انتخاب معیار شباهت هوموفیلی 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 به‌ترتیب در جداول (۶ و ۷) آورده شده است.

برای به‌دست‌آوردن نتایج از معیار ارزیابی AUC که در بخش قبل توضیح داده شد، استفاده شده است. با این تفاوت که به‌جای استفاده از معیار شباهت تجمیع‌شده در آن از

[1] Precision
[2] Recall
[3] FMeasure
[4] Classification Accuracy
[5] Area Under Curve (AUC)

[DOI: 10.29252/jsdp.16.4.45]

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

(جدول-۴): ارزیابی معیار اصلح از بین معیارهای نامزد شباهت هوموفیلی برای مجموعه‌داده شبکهٔ 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	صفت مورد مطالعه	معیار شباهت ساختاری
0.55	جنسیت	OF
0.52	وضعیت تاهل	
0.53	جنسیت	IOF
0.52	وضعیت تاهل	

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 هوموفیلی، این معیار انتخاب شده است.

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

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

میانگین AUCهای به دست آمده	معیار شباهت ساختاری
0.56	شباهت شبکه
0.56	جاکارد
0.53	شباهت کوسینوسی
0.51	نرمال L1

(جدول-۶): ارزیابی معیار اصلح از بین معیارهای نامزد شباهت هوموفیلی برای مجموعه‌داده شبکهٔ اجتماعی مجازی پوکک

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

میانگین AUCهای به دست آمده	صفت مورد مطالعه	معیار شباهت ساختاری
0.49	جنسیت	OF
0.46	جنسیت	IOF

همان‌طور که در جداول (۵ و ۷) شان داده شده است، دو معیار شباهت شبکه و جاکارد، توانستند بیش‌ترین دقت را نسبت به سایر معیارها در هر دو مجموعه‌داده کسب کنند (گفتنی است که اعداد نوشته‌شده همگی به دو رقم اعشار گرد شده‌اند). با توجه به برابرشدن دو معیار شباهت شبکه و جاکارد، تصمیم بر آن شد تا در این پژوهش از معیار شباهت شبکه برای تعیین شباهت ساختاری، بهره گرفته شود.

(جدول-۷): ارزیابی معیار اصلح از بین معیارهای نامزد شباهت ساختاری برای مجموعه‌داده شبکهٔ اجتماعی مجازی پوکک

(Table-7): The evaluation of best similarity metrics between homophily similarity metrics on Pokec Social

میانگین AUCهای به دست آمده	معیار شباهت ساختاری
0.79	شباهت شبکه
0.79	جاکارد
0.78	شباهت کوسینوسی
0.78	نرمال L1

۲-۳-۴- ارزیابی و تحلیل نتایج

در این بخش توسط معیار ارزیابی مطرح شده، به ارزیابی و تحلیل نتایج می‌پردازیم. ویژگی‌های شباهت به دو دسته هوموفیلی و ساختاری تقسیم می‌شوند. دسته ویژگی ساختاری تنها شامل یک ویژگی است؛ اما در دسته ویژگی‌های هوموفیلی، 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:

۱- زمانی که چندین ویژگی کاربران یک شبکه اجتماعی دردسترس باشد، برای استفاده از معیار شباهت هوموفیلی 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 صورت‌گرفته وزن هر یک از ویژگی‌های هوموفیلی 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 به‌دست‌آمده برای ویژگی‌های هوموفیلی به تجمیع وزن‌دار آن‌ها پرداخت و نتایج حاصل آن را با نتایج میانگین‌گیری ساده مقایسه کرد. نتایج حاصل به‌واسطه معیار ارزیابی مطرح‌شده بر روی مجموعه‌داده شبکهٔ اجتماعی مجازی دانشگاه علوم پایه زنجان در جداول (۸ و ۹) ارائه شده است (با توجه به دردسترس‌بودن تنها یک ویژگی هوموفیلی برای مجموعه‌داده شبکه اجتماعی مجازی پوکک، از این مجموعه‌داده در این بخش استفاده نشده است).

[DOI: 10.29252/jsdp.16.4.45]

همان‌طور که در جدول (۹) مشاهده می‌شود، نتایج حاصل بر روی شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان حاکی از آن است که با تخصیص وزن‌ها به ویژگی‌های جنسیت و وضعیت تأهل و استفاده از آن در تجمیع آن‌ها برای به‌دست‌آوردن معیار شباهت، توانسته‌ایم بر دقت این روش نسبت به زمانی که، وزن هر معیار مشابه در نظر گرفته شده است، ۰/۰۲ بیفزاییم. از آنجا که وزن‌های حاصل‌شده اختلاف ناچیزی با یکدیگر دارند، درنظرگرفتن وزن‌ها در تجمیع در قیاس با حالت پایه (درنظرگرفتن وزن یکسان برای هر دو معیار) بهبود اندکی مشاهده شده است. به‌طورطبیعی هر چه اختلاف وزن‌ها برای معیارهای شباهت هوموفیلی بیشتر باشند، اختلاف دقت نتیجه روش ارائه‌شده 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 شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی علوم پایه

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

میانگین AUCها	میانگین وزن‌ها	
0.56	0.48	شباهت ساختاری (NS)
0.55	0.29	شباهت هوموفیلی (جنسیت)
0.52	0.24	شباهت هوموفیلی (وضعیت تاهل)

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

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

میانگین AUCهای به‌دست‌آمده	
0.55	تجمیع ساده دو معیار هوموفیلی
0.57	تجمیع وزن‌دار دو معیار هوموفیلی

۲- محاسبهٔ وزن‌های ویژگی‌های هوموفیلی و ساختاری علاوه‌بر حل چالش بالا (که توضیح داده شد)، می‌تواند در تجمیع ویژگی‌های هوموفیلی و ساختاری کمک کند، که این امر از اهمیت بیشتری برخوردار بوده و هدف اصلی این پژوهش است. راه‌کار پیشنهادی به این شرح است که ابتدا هر یک از ویژگی‌های هوموفیلی و ساختاری با معیار شباهت متناظر خود محاسبه و سپس به‌واسطهٔ وزن‌های حاصل‌شده به‌ازای هر یک از آن‌ها با یکدیگر تجمیع شوند. همان‌طور که عنوان شد در ادبیات موضوع برای مسأله پیش‌بینی یال دو رویکرد وجود دارد (نزدیکی درگراف و هوموفیلی)، با استفاده از راه‌کار 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 شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی علوم پایه.

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

میانگین AUCها	
0.57	تجمیع ساده (جنسیت، وضعیت تاهل و ساختاری)
0.60	تجمیع وزن‌دار (جنسیت، وضعیت تاهل و ساختاری)

در جدول (۱۰) نتایج تجمیع معیارها به صورت وزن‌دار (راه‌کار پیشنهادی) و میانگین‌گیری ساده برای شبکهٔ زنجان، ارائه شده است. روش پیشنهادی توانسته است، نسبت به سایر نتایج (معیار شباهت جنسیت به‌تنهایی، معیار شباهت وضعیت تأهل به‌تنهایی، تجمیع ساده دو معیار یادشده و تجمیع دو معیار یادشده به‌صورت وزن‌دار) دقت را ۰/۰۳ افزایش دهد. همان‌طور که عنوان شد نزدیکی وزن‌های حاصله باعث شده است که میزان افزایش دقت قابل ملاحظه نشود؛ درواقع میزان افزایش دقت مسأله وابسته به میزان تفاوت وزن ویژگی‌های حاصل‌شده (میزان تفاوت اهمیت ویژگی‌های شبکه) است.

همان‌طورکه 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 نشان‌دهندهٔ 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 پیاده‌سازی‌شده برای حالت تصادفی برابر است با مقدار پنجاه درصد، لذا انتظار می رود که (با توجه به این‌که در این شبکه هتروفیلی وجود دارد) برای معیار شباهت هوموفیلی جنسیت این مقدار زیر پنجاه درصد حاصل شود.

[DOI: 10.29252/jsdp.16.4.45]

(جدول-۱۱): نتایج وزن‌های حاصل شده به‌ازای هر ویژگی شباهت برای مجموعه داده پوکک

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

	میانگین وزن‌ها	میانگین AUCها
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، اضافه‌کردن ویژگی هوموفیلی به ساختاری به‌طورطبیعی باید دقت را کاهش دهد. همان‌طور که در جدول (۱۱) مشاهده می‌شود، میزان مساحت زیرنمودار محاسبه‌شده برای ویژگی ساختاری به‌تنهایی برابر ۰/۷۹ است که با اضافه‌شدن معیار هوموفیلی به‌صورت وزن‌دار این عدد به ۰/۶۸ کاهش می یابد. با این حال همان‌طور که در جدول (۱۲) مشاهده می‌شود، اگر اضافه‌کردن معیار هوموفیلی جنسیت به‌صورت میان‌گیری ساده، اضافه شود، مقدار ۰/۷۹ به ۰/۶۵ کاهش می‌یابد. این امر بدان معناست که با احتساب این‌که در شبکه هتروفیلی وجود دارد، ولی میزان اهمیت آن از ویژگی ساختاری کمتر بوده (کمتربودن وزن ویژگی هوموفیلی جنسیت از ویژگی ساختاری در جدول ۱۱) بنابراین در تجمیع وزن‌دار از میزان اهمیت آن نسبت به میانگین‌گیری ساده کاسته شده و توانسته است، نزول کمتری پیدا کند.

(جدول-۱۲): نتایج تجمیع معیارهای هوموفیلی و ساختاری برای مجموعه داده پوکک

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

	میانگین AUCها
تجمیع ساده (معیار هوموفیلی و ساختاری)	**0.65**
تجمیع وزن‌دار (معیار هوموفیلی و ساختاری)	**0.68**

۵- نتیجه‌گیری

در این پژوهش در ابتدا مجموعه‌داده شبکهٔ اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان را جمع‌آوری کردیم. این مجموعه‌داده دارای نقص اطلاعات کاربران بود. برای حل آن از طریق پرسش‌نامه و درنهایت به‌واسطهٔ پیاده‌سازی یکی از 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 به‌عنوان نامزد در نظر گرفتیم و با ارزیابی نتایج روی دو مجموعه‌داده، بهترین معیارها را برگزیدیم.

در این پژوهش توانستیم با ارائه یک مدل به تجمیع وزن‌دار معیارهای شباهت هوموفیلی وساختاری بپردازیم. همچنین به واسطه آن توانستیم به دقت بهترین نتایج معیارهای مبتنی بر همسایگی گره بیافزاییم. 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 به‌علت 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 روش‌هایی وجود دارد که با استفاده از هوموفیلی و نزدیکی در گراف به تشخیص اجتماعات می‌پردازد. با استفاده از نتایج این پژوهش و تخصیص وزن‌های هوموفیلی و نزدیکی در گراف می‌توان به دقت این روش‌ها کمک کرد. در این پژوهش معیاری برای هتروفیلی معرفی نشد که می‌توان به‌عنوان کارهای آینده به آن پرداخت و نتایج این پژوهش را غنی‌تر کرد.

۶- مراجع 6- References

[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.

[DOI: 10.29252/jsdp.16.4.45]

[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. Salehi, S. Shah, M. Cobzarenco, 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.

[۱۰] م. صالحی، ح. عبداللهیان و ع. اسحاق‌پور، "هوموفیلی در شبکه‌های اجتماعی مجازی (مطالعه موردی شبکه اجتماعی مجازی دانشگاه تحصیلات تکمیلی زنجان)،" فصلنامه مطالعات رسانه‌های نوین، جلد ۶، صفحه ۹۱ تا ۱۱۹، ۱۳۹۵.

[10] A. Eshaghpour, 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.

[11] S. Boriah, V. Chandola and V. Kumar, "Similarity measures for categorical data: A comparative evaluation," Red, vol. 30, no. 2, pp. 3, 2008.

[12] C. G. Akcora, B. Carminati and E. Ferrari, "User similarities on social networks," Social Network Analysis and Mining, vol. 3, no. 3, pp. 475-495, 2013.

[۱۳] م. حسینی، م. نصرالهی، ع. بقایی، "یک سامانه توصیه‌گر ترکیبی با استفاده از اعتماد و خوشه‌بندی دوجهته به منظور افزایش کارایی پالایش گروهی"، فصل نامه پردازش علائم و داده‌ها، جلد ۱۵، شماره ۲، صفحه ۱۱۹-۱۳۲، ۱۳۹۷.

[13] Hosseini M, Nasrollahi M, Baghaei A. "A hybrid recommender system using trust and bi-clustering in order to increase the efficiency of collaborative filtering". JSDP.; vol.15 (2) . pp.119-132, 2018.

[14] L. Lü and T. Zhou, "Link prediction in complex networks: A survey," Physica A: 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

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 به‌ترتیب در سال‌های ۱۳۹۲ و ۱۳۹۶ در دانشگاه IASBS و دانشگاه تهران به پایان رسانید. وی هم‌اکنون به‌عنوان کارشناس ارشد SOC با هلدینگ بهین راه‌کار همکاری می‌کند. زمینه‌های پژوهشی مورد علاقه ایشان شبکه‌های اجتماعی مجازی، الگوریتم‌های ژنتیک چندهدفه و امنیت اطلاعات است.

نشانی رایانامه ایشان عبارت است از:

a_eshaghpoor@ut.ac.ir

مصطفی صالحی کارشناسی و کارشناسی ارشد خود را در رشتهٔ مهندسی کامپیوتر به‌ترتیب در سال‌های ۱۳۸۴ و ۱۳۸۶ به پایان رسانید. او در سال ۱۳۹۱ موفق به دریافت درجه دکترا در همین رشته از دانشگاه شریف شد. پس از آن پسادکترا خود را به‌ترتیب در سال‌های ۱۳۹۴ و ۱۳۹۵ در دانشگاه بولونیا کشور ایتالیا و دانشگاه تلکام سود پاریس (فرصت مطالعاتی) گذراند. وی هم اکنون به‌عنوان 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 رایانامه ایشان عبارت است از:

mostafa_salchi@ut.ac.ir

وحید رنجبر کارشناسی و کارشناسی ارشد خود را در رشته مهندسی فناوری اطلاعات به‌ترتیب 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 هم‌اکنون به‌عنوان 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 رایانامه ایشان عبارت است از:

vranjbar@yazd.ac.ir

[DOI: 10.29252/jsdp.16.4.45]