تشخیص‌شبکه‌های چگال از شبکه‌های شبه‌درختی برای پیش‌بینی لینک بر مبنای معیارهای همگنی و ناهمگنی

مهرداد رفیعی‌پور[۱]، زهرا عبدالعلی‌زاده[۱] و سید مهدی وحیدی‌پور[۲،۳]

[۱] دانشجوی کارشناسی ارشد، گروه آموزشی مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان

[۲] استادیار، گروه آموزشی مهندسی کامپیوتر، دانشکده مهندسی برق و کامپیوتر، دانشگاه کاشان، کاشان
[۳] پژوهشکده علوم کامپیوتر، پژوهشگاه دانش‌های بنیادی (IPM)، تهران، ایران.

چکیده

پیش‌بینی لینک، وجود یا عدم وجود ارتباط بین دو موجودیت را بر اساس ویژگی‌های موجودیت‌ها و دیگر لینک‌های مشاهده شده در گراف بررسی می‌کند. الگوریتم‌های پیش‌بینی لینک متفاوتی تا بحال معرفی شده‌اند. این مقاله دو نوع الگوریتم پیش‌بینی لینک را برای شبکه‌های خلوت و چگال بررسی می‌کند؛ در ریاضیات، یک گراف چگال گرافی است که تعداد یال‌های آن نزدیک به بیشینه تعداد یال‌ها باشد و در مقابل یک گراف با کمینه‌ی تعداد یال‌ها یک گراف خلوت است. در این مقاله مقدار ضریب خوشگی برای گراف‌ها با ساختار همگنی و ناهمگنی متفاوت محاسبه شد. سپس بهترین الگوریتم پیش‌بینی لینک برای آن گراف‌ها مشخص شد. در نتیجه، مقداری از ضریب خوشگی بدست آمد که با استفاده از آن می‌توان الگوریتم مناسب برای پیش‌بینی لینک در شبکه را تشخیص داد. به این دلیل که تعداد گراف کافی برای بدست آوردن مقدار مناسب ضریب خوشگی وجود نداشت، روشی را برای تولید گراف تصادفی معرفی کردیم و با استفاده از آن، نقطه مرزی برتری الگوریتم‌های مبتنی بر همسایه مشترک و الگوریتم‌های مبتنی بر درجه نود گراف را بدست آوردیم.

کلمات کلیدی

شبه درختی، همگنی، ناهمگنی، گراف، شبکه، پیش‌بینی لینک، بارابا‌سی–البرت، ضریب خوشگی، همسایه مشترک

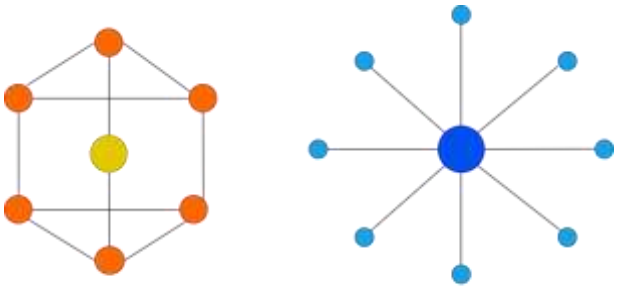
۱- مقدمه

پیش‌بینی لینک در سال‌های اخیر مورد توجه قرار گرفته است. الگوریتم‌های پیش‌بینی لینک برای مثال می‌توانند در شناسایی ارتباطات دیده‌نشده نقش داشته تا هزینه آزمایشات را کاهش دهند [۱]. از طرف دیگر پیش‌بینی لینک در طراحی الگوریتم سیستم‌های تصمیم‌یار کاربرد دارد [۲]. پیش‌بینی لینک، وجود یا عدم وجود ارتباط بین دو موجودیت را بر اساس ویژگی‌های موجودیت‌ها و دیگر لینک‌های مشاهده شده در گراف بررسی می‌کند. گاهی لینک جدیدی که تا به حال مشاهده نشده را بدست می‌آورد و گاهی لینک جعلی را حذف می‌کند.

برخی روش‌های پیش‌بینی لینک با افزایش تعداد لینک در شبکه، عملکرد بهتری خواهند داشت. به عنوان مثال روش‌های پیش‌بینی لینک که بر اساس همسایه مشترک عمل می‌کنند. احتمال ایجاد لینک میان دو نود که همسایه مشترک دارند بسیار بالاست. به عنوان مثال فرض کنید که در یک شبکه، نودها نشان‌دهنده‌ی اشخاص و لینک میان دو نود نشان‌دهنده‌ی دوستی میان دو نفر باشد. حال اگر دو نفر دوست مشترک داشته باشند، احتمال آنکه در آینده با هم دوست شوند بسیار بالاست. اما لازمه استفاده از چنین روشی وجود تعداد مناسب لینک در شبکه است.

در ریاضیات و بر اساس تعداد لینک‌ها در گراف، گراف چگال[۱] و گراف خلوت یا شبه‌درختی[۲] مطرح می‌شود. یک گراف چگال گرافی است که تعداد یال‌های آن نزدیک به بیشینه تعداد یال‌ها باشد و در مقابل یک گراف با کمینه تعداد یال‌ها یک گراف خلوت یا شبه درخت است. با توجه به این تعاریف، برخی روش‌های پیش‌بینی لینک که در شبکه‌های چگال و شلوغ عملکرد مناسبی دارند، ممکن است در گراف‌های شبه درختی به خوبی عمل نکنند. به خصوص روشهایی که معیار در آن‌ها وجود همسایه مشترک (و تشکیل شکل مثلث توسط سه یال) در گراف است. هنگامی که گراف شبه درختی باشد، تعداد مثلث‌های تشکیل شده در گراف معیار مناسبی برای تصمیم‌گیری در خصوص پیش‌بینی لینک نیست.

برای پیش‌بینی لینک در شبکه‌های خلوت معیار جدیدی بر اساس ساختار کلی شبکه ارائه شده است [۳]. این معیار بر اساس ناهمگنی و همگنی[۳] شبکه، لینک را پیش‌بینی می‌کند. در شبکه‌ی همگن نودهای با اختلاف درجه‌ی کم و در شبکه ناهمگن نودهای با اختلاف درجه زیاد به هم لینک دارند. بدین صورت در یک شبکه همگن نودهایی که به یکدیگر متصل هستند، درجه مشابه دارند. و به طور مقابل در شبکه ناهمگن نودهایی که اختلاف درجه زیادی دارند به هم متصل‌اند. در شکل (۱) گراف همگن و ناهمگن با یک مثال نشان داده شده‌اند.



شکل ۱: در گراف(الف) نودها با اختلاف درجه زیاد و در گراف(ب) نودها با اختلاف درجه کم به هم متصل‌اند.

در این مقاله یک سوال اصلی وجود دارد: برای پیش‌بینی لینک، چه زمانی استفاده از معیارهای مبتنی بر همسایه مشترک و چه زمانی استفاده از معیار مبتنی بر ناهمگنی کارایی بیشتری دارد؟ در این مقاله برای پاسخ به این سوال، استفاده از ضریب خوشگی[۴] پیشنهاد شده است. ضریب خوشگی نسبت تعداد مثلث‌های بسته به مجموع تعداد مثلث‌های باز و بسته در گراف است. نتیجه استفاده از این ضریب پیشنهاد یک مرز برای مقادیر مختلف ضریب خوشگی است. چنانچه در بالا و پایین این مرز پیشنهاد مشخصی برای روش پیش‌بینی لینک ارایه می‌شود: در پایین این مرز یا مقدار، استفاده از معیار ناهمگنی و در محدوده بالای آن استفاده از معیار همسایگی مشترک مناسب است.

همچنین در این مقاله، مدل جدیدی برای ساخت شبکه پیشنهاد می‌شود که در آن شبکه‌هایی با ضریب خوشگی متفاوت ساخته می‌شود. با استفاده از این مدل محدوده‌های پیشنهادی را بررسی کرده‌ایم؛ برای بررسی دقیق نیاز به چندین شبکه متفاوت داریم ولی تعداد شبکه‌های خلوت واقعی بسیار کم است. در مدل پیشنهادی با تنظیم پارامتر مدل شبکه‌های متفاوتی برای تشخیص مرز استفاده از معیارها تولید می‌شود. بر روی شبکه‌های ساخته شده روش‌های متفاوت پیش‌بینی لینک مقایسه می‌شود تا درستی محدوده‌های پیشنهادی بررسی شود. نهایتاً محدوده‌های بدست آمده بر روی ضریب خوشگی بر روی چند شبکه واقعی نیز آزمایش شده است. در آزمایش‌های انجام شده، تحلیل و مقایسه روش‌های پیش‌بینی لینک بر اساس معیار AUC است که یکی از معیارهای رایج در این حوزه است [۴].

در ادامه بخش‌های زیر به ترتیب ارایه می‌شوند. در بخش دوم کارهای پیشین شرح داده می‌شوند. در بخش سوم ادبیات پژوهش مرور و در بخش چهارم مسئله توضیح داده می‌شود. بخش پنجم روش پیشنهادی را توضیح می‌دهد. بخش ششم آزمایشات روی چند مجموعه داده و بخش هفتم نتیجه‌گیری را ارائه می‌کند.

۲- کارهای پیشین

تا به حال معیارها و الگوریتم‌های متفاوتی برای پیش‌بینی لینک ارائه شده‌اند. برخی از آن‌ها در دسته روش‌های مبتنی بر شباهت دسته‌بندی می‌شوند. شباهت دو نود ممکن است وابسته به تعداد زیادی از ویژگی‌های مشترک آن نودها باشد. از آن‌جایی که این ویژگی‌ها معمولا پنهان هستند از معیارهای شباهت ساختاری[۵] که مبتنی بر ساختار شبکه هستند استفاده می‌شود. این معیارها با توجه به ساختار شبکه، ممکن است برای یک نوع شبکه به خوبی عمل کنند اما برای شبکه دیگر نه. یک نوع از معیارهای شباهت ساختاری شاخص‌های محلی[۶] هستند که وابسته به اطلاعات محلی‌اند. از جمله معروف‌ترین معیارهای محلی می‌توان به معیار ژاکارد[۷] [۵] و آدامیک‌آدار[۸] [۶] اشاره کرد. هم‌چنین معیار کتز[۹] یک معیار شباهت کلی[۱۰] است که وابسته به ساختار کلی شبکه و همسایه مشترک است، این دسته از معیارها پیچیدگی محاسباتی بالایی دارند [۳] و [۸]. الگوریتم‌های سنتی پیش‌بینی لینک از این معیارهای محلی و کلی استفاده می‌کنند؛ در بیشتر آن‌ها نقش همسایه مشترک بسیار پررنگ است به صورتی که احتمال ایجاد لینک میان دو گره که همسایه مشترکی دارند، بالاتر می‌رود.

از آن‌جایی که بسیاری از شبکه‌ها در دنیای واقعی تعداد یال‌های کمی دارند و خلوت هستند، با توجه به تکیه‌ی روش‌های سنتی بر همسایه مشترک، کارایی این روش‌ها در شبکه‌های خلوت پایین‌تر است. بدین ترتیب معیار پیش‌بینی لینک جدیدی برای شبکه‌های شبه درختی ارائه شد [۳]. این معیار بر اساس اختلاف درجه نودها به میزان شباهت بین هر دو نود یک امتیاز می‌دهد. چون این معیار مبتنی بر اختلاف درجه نودهای گراف است؛ بنابراین معیار همگنی و ناهمگنی را در در هر نوع شبکه، چگال و خلوت می‌توان استفاده نمود، اما استفاده از معیارهای مبتنی بر همسایه مشترک، مختص به شبکه‌های چگال است.

۳- ادبیات پژوهش

در این بخش تعاریف مورد نیاز به ترتیب ارایه می‌شوند.

۳-۱- پیش‌بینی لینک

پیش‌بینی وجود ارتباط بین دو موجودیت بر اساس ویژگی‌های موجودیت‌ها و دیگر لینک‌های مشاهده شده در گراف را پیش‌بینی لینک می‌گویند.

۳-۲- ضریب خوشگی

ضریب خوشگی بر پایه یک سه تایی از گره‌ها تعریف می‌شود. یک سه تایی متشکل از سه گره متصل به هم. بنابراین یک مثلث شامل سه سه‌تایی است که هر یک به مرکزیت یکی از گره‌هاست. ضریب خوشگی نسبت تعداد کل سه‌تایی‌های بسته (یا سه برابر تعداد کل مثلث‌ها) به تعداد کل سه‌تایی‌هاست (سه‌تایی‌های باز و بسته). ضریب خوشگی در معادله (۱) تعریف شده است (C اندازه ضریب خوشگی)

$$C = \frac{\text{تعداد سه تایی های بسته}}{\text{تعداد سه تایی های بسته و باز}} \qquad (۱)$$
$$0 \le C \le 1$$

۳-۳- معیار AUC

گراف $G = (V, E)$ توسط مجموعه رئوس V و مجموعه یال‌های E توصیف می‌شود. اعضای مجموعه یال‌ها، جفت‌های بدون ترتیب از مجموعه رئوس هستند: $e = (v_i, v_j) \in E$ که $v_i, v_j \in V$. جفت (v_i, v_j) حداکثر در یک یال e خواهد بود که $e \in E$. مسئله پیش‌بینی لینک استاندارد بدین صورت فرموله می‌شود. مجموعه یال‌های E به دو بخش تقسیم می‌شود: E^T و E^P که $E^T \cup E^P = E$ و $E^T \cap E^P = \emptyset$. تقسیم کردن مجموعه E به E^P (که شامل ۱۰٪ لینک‌های مشاهده شده است.) و E^T (که شامل ۹۰٪ لینک‌های مشاهده شده است.) قراردادی است و از آن برای محاسبات استفاده می‌شود. همه لینک‌های $E^T \cup E^P = E$ مشاهده شده هستند، لینک‌های موجود در E^T مجموعه آموزشی را تشکیل می‌دهند و برای پیاده سازی یک الگوریتم پیش‌بینی لینک استفاده می‌شود و نتیجه آن روی مجموعه E^P ارزیابی خواهد شد.

ناحیه زیر منحنی (AUC) تا به حال به طور گسترده برای اندازه گیری دقت پیش‌بینی استفاده شده است [۴]. در این مقاله برای اندازه گیری دقت الگوریتم‌های پیش‌بینی لینک از AUC استفاده می‌کنیم. تنها از دانش E^T برای محاسبه اندازه شباهت $\text{Score}_{x,y}$ در الگوریتم‌ها اجازه استفاده هست. ابتدا شباهت m جفت نود تصادفی از E^P و E' را محاسبه می‌کنیم. اگر m' بار شباهت محاسبه شده در E^P بزرگتر از شباهت محاسبه شده در E' باشد و

m'' بار هر دو شباهت حساب شده مساوی باشند، آنگاه AUC با استفاده از معادله (۲) بدست می‌آید [۹]:

$$AUC = (m' + 0.5m'')/m \qquad (۲)$$

هرچه AUC به ۱ نزدیک‌تر باشد، الگوریتم پیش‌بینی لینک عملکرد بهتری دارد.

۳-٤- معیار پیش‌بینی لینک مبتنی بر ناهمگنی و همگنی

معیارهای شباهتی که برای شبکه‌های شبه درختی (یا شبکه‌هایی با لینک‌های بسیار کم) پیشنهاد شده معیار شباهت همگنی و معیار شباهت ناهمگنی هستند.

معیار شباهت ناهمگنی (HEI)

این معیار با نماد s_{ij}^{HEI} در معادله (۳) بیان شده است که در آن $k(i)$ درجه نود i و $k(j)$ درجه نود j و α توان آزاد ناهمگنی است. برای شبکه‌هایی با ناهمگنی بیشتر و همگنی کمتر معرفی می‌شود؛ هر چه اختلاف درجه دو نود بیشتر باشد s_{ij}^{HEI} بزرگتر است [۳].

$$s_{ij}^{HEI} = |k(i) - k(j)|^{\alpha} \qquad (۳)$$

معیار شباهت همگنی (HOI)

این معیار برای شبکه‌هایی با همگنی بیشتر و ناهمگنی کمتر ارایه شده است. هر چه اختلاف درجه دو نود کمتر باشد s_{ij}^{HOI} بزرگتر است. در معادله (۴)، $k(i)$ درجه نود i و $k(j)$ درجه نود j و α توان آزاد همگنی است [۳].

$$s_{ij}^{HOI} = \frac{1}{|k(i) - k(j)|^{\alpha}} \qquad (۴)$$
$$0 \le \alpha \le 1$$

۳-٥- معیار سنتی پیش‌بینی لینک

معیار ژاکارد قدیمی‌ترین معیار پیش‌بینی لینک است و بر اساس نقش همسایه-های مشترک می‌باشد [۵] و [۸]. معادله (۶) اندازه معیار ژاکارد میان نود i و نود j را با نماد $s_{ij}^{Jaccard}$ نشان می‌دهد که در آن $\Gamma(i)$ مجموعه همسایه‌های نود i و $\Gamma(\mathrm{i,j})$ مجموعه همسایه‌های مشترک دو نود i و j هستند.

$$s_{ij}^{Jaccard} = \frac{|\Gamma(\mathrm{i,j})|}{|\Gamma(\mathrm{i}) \cup \Gamma(\mathrm{j})|} \qquad (۵)$$

۳-٦- مدل باراباسی-البرت[۱۱]

مدل باراباسی-البرت یک الگوریتم تولید ایجاد شبکه پیچیده مستقل از مقیاس با مکانیزم وابستگی امتیازی است. در این شبکه گره‌هایی با درجه بالای غیر عادی در مقایسه با سایر گره‌های شبکه تولید می‌شوند. برای تولید گراف بر اساس این مدل الگوریتم باراباسی-البرت بدین شکل عمل می‌کند: شبکه با یک گراف اولیه همبند شامل m_0 نود شروع می‌کند. هر بار یک نود به گراف اضافه میشود و هر نود جدید با احتمال p_i (فرمول (۶)) به $m \le m_0$ نودهای قبلی وصل می شود، که:

$$p_i = \frac{k_i}{\sum_j k_j} \qquad (۶)$$

به طوری که k_i درجه نود i است [۱۰].

٤- مسأله

با معرفی معیارهای HEI و HOI مبتنی بر اختلاف درجات نود و وجود معیارهای سنتی مبتنی بر همسایه مشترک، نمی‌توان عنوان کرد چه زمانی از کدام معیار برای پیش‌بینی لینک استفاده شود. در واقع تعدد معیارهای پیش‌بینی لینک و عدم شناخت کافی نسبت به ساختار شبکه، باعث تردید در انتخاب معیار مناسب می‌شود. پس مسئله اصلی که در این مقاله به دنبال پاسخ به آن هستیم این است: برای پیش‌بینی لینک، چه زمانی استفاده از معیارهای مبتنی بر همسایه مشترک و چه زمانی استفاده از معیارهای HEI و HOI کارایی بیشتری دارد؟

٥- راه حل پیشنهادی

در این مقاله برای حل مسئله فوق استفاده از ضریب خوشگی پیشنهاد می‌شود. یعنی با توجه به مقدار ضریب خوشگی در شبکه تعیین می‌شود که از کدام معیار برای پیش‌بینی لینک استفاده شود. از آنجایی که اندازه ضریب خوشگی وابسته به تعداد مثلث‌های باز و بسته یعنی همان نقش همسایه مشترک است؛ بنابراین عملکرد معیارهای مبتنی بر همسایه مشترک با اندازه ضریب خوشگی در ارتباط است. بنابراین ضریب خوشگی بالاتر نشان دهنده وجود تعداد بالاتر همسایه مشترک در شبکه است که برای معیارهای سنتی مناسب‌تر هستند.

برای بررسی دقیق روش پیشنهادی لازم است تا آن را بر روی شبکه‌های مختلفی از انواع مختلف شبه‌درختی و چگال امتحان کنیم. اما مشکل آن است که نمونه‌هایی از انواع مختلف شبکه‌ها در دست نیست. برای حل این مشکل، در ادامه یک مدل تصادفی ساخت شبکه با نام مدل R ارایه می‌شود. این مدل مبتنی بر مدل باراباسی-آلبرت پیشنهاد شده است که در آن با تنظیم پارامتر d ضریب خوشگی در شبکه ساخته شده متفاوت خواهد بود؛ پارامتر d احتمال تولید مثلث در شبکه است و با کاهش آن احتمال تولید شبکه شبه‌درختی افزایش می‌یابد. در این مقاله، مدل R برای ساخت شبکه ناهمگن توسعه داده شده است.

٥-١- مدل R

در این مدل یک گراف اولیه، یعنی $G = (V, E)$، که توسط مدل باراباسی-البرت تولید شده، دریافت می‌گردد . نود $i \in V$ را در این گراف در نظر بگیرید. این نود می‌تواند با کمک یکی از همسایگان خود، مثلا $z \in V$، و با احتمال d یک مثلث ایجاد کند. بدیهی است با افزایش مقدار d تعداد مثلث‌های ایجاد شده در گراف اولیه افزایش می‌یابد. نود i برای ساخت مثلث باید یکی از همسایگان نود z را انتخاب کرده و به آن وصل شود. از آنجایی که مدل R برای تولید شبکه‌های ناهمگن توسعه داده شده است، نود i تمایل دارد تا نودی را انتخاب کند که بیشترین تفاوت را در درجه با خودش دارد. از آنجایی که مدل R یک مدل تصادفی است انتخاب نود توسط نود i با احتمال

بیان می‌شود؛ احتمال اتصال نود i به نودی که اختلاف درجه بیشتری با آن دارد بالاتر است.

فرض کنید نود $j \in \Gamma(z)$ یکی از همسایگان نود z باشد. احتمال انتخاب شدن نود j توسط نود i را معادله (۷) با $P_{select}(i,j)$ نشان داده است. در این معادله، $K(i)$ درجه نود i را مشخص می‌کند و $\Gamma(z)$ مجموعه همسایگان نود z را مشخص می‌کند. شبه‌کد (۱) مدل R پیشنهادی را نشان می‌دهد. دراین شبه‌کد، خط ۴ با احتمال d اجرا می‌شود. در واقع با افزایش احتمال d امکان ایجاد مثلث افزایش می‌یابد.

$$P_{select}(i,j) = \frac{|K(i)-K(j)|}{\sum_{r\in\Gamma(z)} |K(i)-K(r)|} \quad , i,j \in \Gamma(z) \qquad (۷)$$

۱. یک گراف تصادفی بر مبنای مدل بارابَاسی-آلبرت بساز

۲. به ازای هر i که $i \in V$:

۳. به ازای هر z درون همسایه‌های i:

۴. به احتمال d از میان j همسایه های z یک یال میان i و j متناسب با

$$P_{select}(i,j) = \frac{|K(i)-K(j)|}{\sum_{r\in\Gamma(z)} |K(i)-K(r)|} \quad , i,j \in \Gamma(z)$$

متصل کن.

شبه‌کد ۱: کد مدل R برای تولید گراف تصادفی ناهمگن با احتمال d

٦- آزمایش‌ها

در این بخش چهار آزمایش مختلف با هدف بررسی رابطه اندازه ضریب خوشگی با ویژگی‌های شبکه و عملکرد معیارهای مبتنی بر همگنی و ناهمگنی و معیار مبتنی بر همسایه مشترک بر اساس معیار AUC طراحی شده است. در این مجموعه آزمایشات برای تولید گراف اولیه بارابَاسی $m = 1$ و تعداد نودهای گراف متغیر است. همچنین برای مدل R، $\alpha = 0.5$ در نظر گرفته شده است.

در این مقاله برای مقایسه و بررسی کارایی معیارها از معیار ژاکارد به عنوان معیار مبتنی بر همسایه مشترک استفاده می‌شود.

٦-١- بررسی رابطه ضریب خوشگی با تعداد نودها

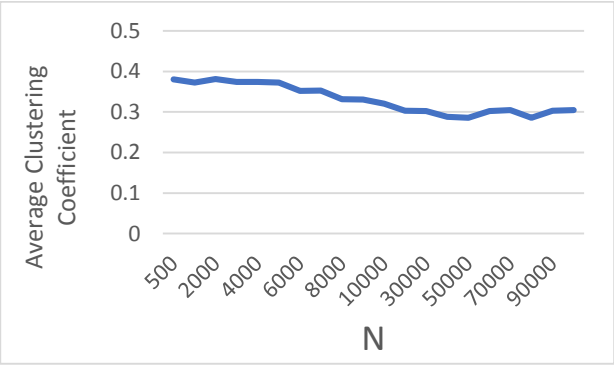
برای گراف‌های تولید شده با مدل R میانگین اندازه ضریب خوشگی در گراف‌هایی با تعداد نودهای مختلف محاسبه و در نمودار شکل (۱) نشان داده شده است. در این سری از آزمایش‌ها اندازه d بین ۰,۲۲۵ تا ۰,۳ متغیر است. همانطور که در نمودار شکل (۱) مشخص هست، اندازه ضریب خوشگی به تعداد نودها (N) وابسته است و با زیاد شدن تعداد نودها ضریب خوشگی به تدریج کاهش پیدا می‌کند. در واقع انتخاب معیار پیش‌بینی لینک به تعداد نودهای یک گراف بستگی دارد.

٦-٢- بررسی اثر d بر ضریب خوشگی

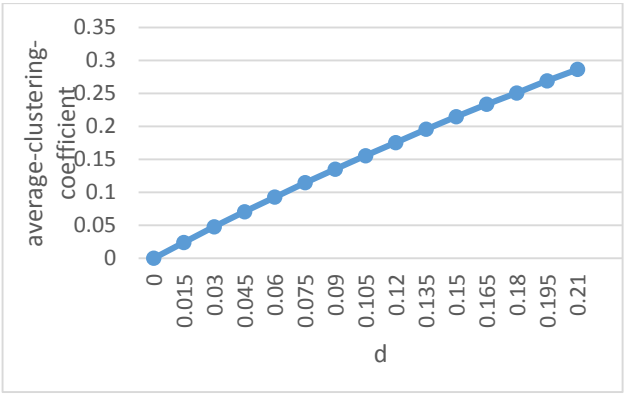
این بخش از آزمایش به بررسی رابطه بین ضریب خوشگی و d می‌پردازد. نتایج این آزمایش در شکل (۲) نمایش داده شده، در نمودار شکل (۲) برای گراف با تعداد نود ثابت $N = 100000$، مشاهده می‌شود که با افزایش احتمال d در مدل تولید شبکه R یعنی افزایش احتمال ایجاد مثلث بسته به طوری که گراف همچنان ناهمگن بماند، همانطور که انتظار می‌رود ضریب خوشگی افزایش یافته است. همچنین برای $d = 0$ ضریب خوشگی صفر است چون احتمال ساختن سه‌تایی بسته صفر است و با افزایش این احتمال، سه‌تایی‌های بسته و در نتیجه ضریب خوشگی افزایش یافته است.

٦-٣- مقایسه عملکرد الگوریتم‌های سنتی و جدید با توجه به تغییرات ضریب خوشگی

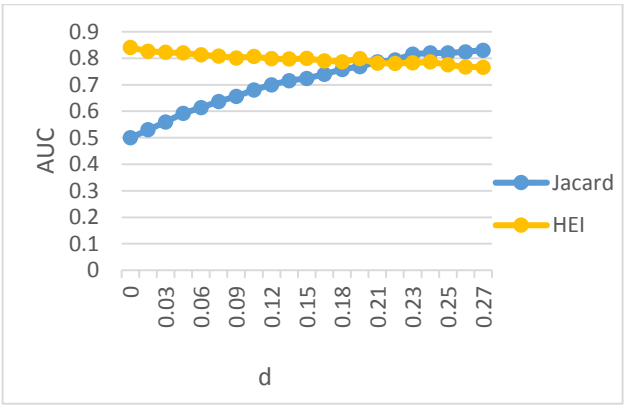
در نمودار شکل (۲) مشخص است که با افزایش احتمال تشکیل سه‌تایی بسته ضریب خوشگی روندی صعودی دارد. پس از آن در آزمایش دیگری رفتار دو معیار مبتنی بر ناهمگنی و معیار مبتنی بر همسایه مشترک با در مقادیر متفاوت d مقایسه شده است. در شکل (۳) نتایج این آزمایش به تصویر کشده شده است، با توجه به نمودار شکل (۳) مشاهده می‌شود که با افزایش احتمال d و بزرگ شدن اندازه ضریب خوشگی ضمن حفظ ناهمگنی گراف شبکه عملکرد معیار ناهمگنی تغییرات آنچنانی ندارد، در حالیکه با چگال شدن گراف رفته رفته عملکرد معیار ژاکارد بهبود چشمگیری پیدا می‌کند. حتی در احتمال $d = 0.21$ از معیار ناهمگنی پیشی می‌گیرد و مقدار AUC بهتری به دست می‌دهد. در نتیجه می‌توان اینگونه استنباط کرد، در گراف‌های خلوت که ضریب خوشگی آنها کمتر از حدود ۰,۳ باشد الگوریتم‌های مبتنی بر ناهمگنی دقت بالاتری در پیش‌بینی لینک دارند و عملکرد الگوریتم مبتنی بر ژاکارد، حدود ۰,۵ است که نتیجه آن بسیار مشابه روش انتخاب تصادفی است [۷]. در گراف‌هایی که ضریب خوشگی بین ۰,۳ و ۰,۴ بوده اختلاف دقت الگوریتم‌ها کمتر شده و نتایج نزدیک به هم دارند. در نتیجه با چگال شدن گراف و افزایش سه‌تایی‌های بسته دیگر الگوریتم‌های مبتنی بر معیار جدید نسبت به الگوریتم مبتنی بر ژاکارد برتری ندارند.



شکل ۲: تغییرات ضریب خوشگی با تغییر تعداد نودهای شبکه (N)



شکل ۳: تغییرات ضریب خوشگی با افزایش *d* در مدل *R* برای شبکه با *N*=100000



شکل ٤: تغییرات *AUC* دو معیار *HEI* و *Jaccard* با افزایش d در مدل R برای شبکه با *N*=100000

٦-٤- مقایسه عملکرد الگوریتم‌های جدید و سنتی با در نظر گرفتن تغییرات ضریب خوشگی روی چند مجموعه داده

مقدار ضریب خوشگی و AUC برای چند مجموعه داده محاسبه شده است. مشاهده می‌شود که برای ضریب خوشگی کمتر از ۰٫۳ معیار مبتنی بر همگنی و ناهمگنی به مراتب بهتر عمل کرده و عملکرد معیار سنتی ژاکارد ضعیف است. اما با نزدیک شدن به ضریب خوشگی ۰٫۳ رفته رفته عملکرد معیار ژاکارد بهتر شده و اختلاف AUC آن با معیارهای جدید کمتر شده است.

همینطور با گذشتن اندازه ضریب خوشگی از مقدار ۰٫۳ عملکرد معیارهای جدید با توجه با زیاد شدن سه تایی‌های بسته روند رو به پایینی دارد و ضعیف‌تر عمل کرده و معیار سنتی ژاکارد که بر پایه همسایه مشترک پیش‌بینی لینک می‌کند، عملکرد بسیار خوبی داشته است و AUC بالا و نزدیک به ۱ بدست آورده است.

۷- نتیجه‌گیری

اندازه ضریب خوشگی با افزایش تعداد نودها و احتمال تشکیل مثلث‌های بسته افزایش میابد. در مسئله پیش‌بینی لینک برای شبکه‌های شبه درختی به دنبال معیاری بودیم که بتوان تشخیص داد الگوریتم‌های مبتنی بر ناهمگنی در چه شرایطی نسبت به الگوریتم‌های سنتی برتری دارند. معیاری که از آن استفاده شد ضریب خوشگی است. نتایج آزمایشات نشان داد با افزایش ضریب خوشگی عملکرد الگوریتم‌های سنتی بهبود می یابد و با نزدیک شدن به ضریب خوشگی ۰٫۳ عملکرد الگوریتم مبتنی بر ژاکارد از الگوریتم‌های جدید پیشی می‌گیرد. بنابراین می‌توان ضریب خوشگی را یک متر برای انتخاب الگوریتم مناسب پیش‌بینی لینک دانست و یک مرز برای استفاده از معیارهای متفاوت مشخص شد.

جدول ۱: مقادیر ضریب خوشگی میانگین و AUC الگوریتم‌های مبتنی بر همگنی و ناهمگنی و الگوریتم ژاکارد برای چند مجموعه داده واقعی. در این جدول Average-Clustering-Coefficient با خلاصه ACC نشان داده شده است.

Data set	ACC	Jaccard AUC	HEI AUC	HOI AUC
Gnutella deer to deer network	۰٫۰۰۵۴	۰٫۵۱۸۴	**۰٫٦٦٥۳**	۰٫۳۳۴۷
SNAD/higgs-twitter	۰٫۰۱۵۶	۰٫۵۳۲۵	**۰٫۹٤۱٦**	۰٫۰۵۸۴
Road network of California	۰٫۰۴۶۳	۰٫۵۶۲	۰٫۴۲۹۳	**۰٫۵۷۰۷**
Road network of Dennsylvania	۰٫۰۴۶۴	۰٫۵۶۴۵	۰٫۴۲۷۸	**۰٫۵۷۲۲**
Road network of Texas	۰٫۰۴۷۰	۰٫۵۵۶	۰٫۴۳۳۹	**۰٫۵٦٦۱**
Slashdot social network	۰٫۰۶۰۳	۰٫۷۳۷۵	**۰٫۹۰۳۳**	۰٫۰۹۶۷
Who-trusts-whom network of Edinions.com	۰٫۱۳۷۷	۰٫۸۷۹۸	**۰٫۸۹۹۱**	۰٫۱۰۰۹
Github-social	۰٫۱۶۷۵	۰٫۸۰۳۸	**۰٫۸۵٦۸**	۰٫۱۴۳۲
Amazon droduct co-durchasing	۰٫۴۱۷۶	**۰٫۹٤۵۵**	۰٫۶۳۶۳	۰٫۳۶۳۷
Ego-facebook	۰٫۶۰۵۵	**۰٫۹۹۱۵**	۰٫۵۰۵۴	۰٫۴۹۴۶

سپاسگزاری

بدینوسیله از حمایت پژوهشگاه دانش‌های بنیادی (IPM) در انجام این تحقیق کمال تشکر و قدردانی را داریم. (شماره قرارداد CS1398-4-207)

مراجع

[1] Lü, Linyuan, et al. "*Toward link predictability of complex networks*", Proceedings of the National Academy of Sciences 112.8, pp. 2325-2330, 2015.

[2] Zhang, Zi-Ke, et al. "*Solving the cold-start problem in recommender systems with social tags*", EPL (Europhysics Letters), vol. 92, no. 2, 2010.

[3] Shang Ke-ke, Li Tong-chen, Small Michael, Burton David, Wang Yan, "*Link prediction for tree-like networks*" chaos, vol. 29, no. 6, p. 061103, 2019.

[4] J. A. Hanley and B. J.McNeil, "The meaning and use of the area under a receiver oderating characteristic (ROC) curve", Radiology, vol. 143, no. 1, pp.29-36, 1982.

[5] D.Jaccard, "*Etude comdarative de la distribution floraledanseune dortion des aldes et des jura*", Bull. Soc. Vaudoise Sci. Nat., vol. 37, pp. 547–579, 1901.

[6] L. A. Adamic and E.Adar, "*friends and neighbors on the web*," Soc. Netw., vol. 25, pp. 211-230, 2003.

[7] F. Fouss, A. Pirotte, J. Renders, and M. Saerens, "*Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation*," IEEE Trans. Knowl. Data. Eng., vol. 19, pp. 355–369, 2007.
[8] L. L¨u, T. Zhou, "*Link prediction in complex networks: A survey*", Physica A: Statistical Mechanics and its Applications, vol. 390, no. 6, pp. 1150–1170, 2011.
[9] Chen, Bolun, et al.,"*Link prediction on directed networks based on AUC optimization*.", IEEE Access 6, pp. 28122-28136, 2018.
[10] Barabási et al., "*Modeling the Internet's large-scale topology*", Proceedings of the National Academy of Sciences,, vol. 99, no. 21, pp. 13382-13386, 2002.

پانویس ها

[1] Dense graph
[2] Sparse graph
[3] Heterogeneity and homogeneity
[4] Clustering-coefficient
[5] Structural similarity
[6] Local Similarity Indices
[7] Jaccard
[8] Adamic-Adar
[9] Katz
[10] Global Similarity Indices
[11] Barabási–Albert