# نمونه‌گیری از گراف شبکه‌های اجتماعی براساس ویژگی‌های توپولوژیکی و الگوریتم کلونی زنبور عسل

**عسگرعلی بویر*[1] و سمیه نوروزی[2]**

[1]گروه مهندسی کامپیوتر، دانشکده فناوری اطلاعات، دانشگاه شهید مدنی آذربایجان، تبریز، ایران

[2]واحد میاندوآب، دانشگاه آزاد اسلامی، میاندوآب، ایران

## چکیده

**با توجه به رشد سریع شبکه‌های اجتماعی در چند سال اخیر، مسأله نمونه‌گیری از گراف‌های بسیار بزرگ شبکه‌های اجتماعی با هدف تجزیه و تحلیل سریع شبکه بر اساس نمونه‌های کوچک، اهمیت خاصی پیدا کرده است. مطالعات زیادی در این راستا انجام شده است، ولی آنها تا حد زیادی با مشکل انتخاب تصادفی، عدم حفظ ویژگی‌های شبکه‌های پیچیده در گراف حاصل و یا صرف هزینه زمانی بالا برای استخراج گراف نمونه مواجه هستند. در این مقاله یک روش نمونه‌گیری جدید را برای نخستین‌بار با ارائه یک رابطه جدید مبتنی بر ویژگی‌های ساختاری برای مشخص‌کردن اهمیت گره‌ها و استفاده از الگوریتم کلونی زنبور عسل پیشنهاد می‌کنیم. این روش نمونه‌گیری با ارائه یک رویکرد آگاهانه غیرتصادفی در نمونه‌گیری سعی دارد تا نمونه حاصله از لحاظ ویژگی‌هایی مانند توپولوژی شبکه، توزیع درجه، تراکم داخلی، درجه ورودی و خروجی و غیره شباهت زیادی با شبکه اصلی داشته باشد. نتایج حاصل، برتری روش پیشنهادی را از لحاظ حفظ ویژگی‌های توزیع درجه، ضریب خوشه‌بندی و غیره در نمونه گراف به‌دست‌آمده نشان می‌دهد.**

**واژگان کلیدی: نمونه‌گیری، شبکه‌های اجتماعی، ضریب خوشه‌بندی، کلونی زنبور عسل**

# Sampling from social networks's graph based on topological properties and bee colony algorithm

**Asgarali Bouyer*[1] & Somayeh Norouzi[2]**

[1]Department of Computer Engineering, Azarbaijan Shahid Madani University, Tabriz, Iran

[2]Department of Computer, Miyandoab branch, Islamic Azad University, Miyandoab, Iran

## Abstract

**In recent years, the sampling problem in massive graphs of social networks has attracted much attention for fast analyzing a small and good sample instead of a huge network. Many algorithms have been proposed for sampling of social network' graph. The purpose of these algorithms is to create a sample that is approximately similar to the original network's graph in terms of properties such as degree distribution, clustering coefficient, internal density and community structures, etc. There are various sampling methods such as random walk-based methods, methods based on the shortest path, graph partitioning-based algorithms, and etc. Each group of methods has its own pros and cones. The main drawback of these methods is the lack of attention to the high time complexity in making the sample graph and the quality of the obtained sample graph. In this paper, we propose a new sampling method by proposing a new equation based on the structural properties of social networks and combining it with bee colony algorithm. This sampling method uses an informed and non-random approach so that the generated samples are similar to the original network in terms of features such as network topological properties, degree distribution, internal density, and preserving the clustering coefficient and community structures. Due to the random nature of initial population generation in meta-heuristic sampling methods such as genetic algorithms and other evolutionary algorithms, in our proposed method, the idea of consciously selecting nodes in producing**

* Corresponding author

*نویسندهٔ عهده‌دار مکاتبات

• تاریخ ارسال مقاله: ۱۳۹۸/۰۲/۱۲ • تاریخ پذیرش: ۱۳۹۸/۱۱/۰۲ • تاریخ انتشار: ۱۳۹۹/۰۹/۱۵ • نوع مطالعه: بنیادی





[ DOI: 10.29252/jsdp.17.3.55 ]

**the initial solutions is presented. In this method, based on the finding hub and semi-hub nodes as well as other important nodes such as core nodes, it is tried to maintain the presence of these important nodes in producing the initial solutions and the obtained samples as much as possible. This leads to obtain a high-quality final sample which is close to the quality of the main network. In this method, the obtained sample graph is well compatible with the main network and can preserve the main characteristics of the original network such as topology, the number of communities, and the large component of the original graph as much as possible in sample network. Non-random and conscious selection of nodes and their involvement in the initial steps of sample extraction have two important advantages in the proposed method. The first advantage is the stability of the new method in extracting high quality samples in each time. In other words, despite the random behavior of the bee algorithm, the obtained samples in the final phase mostly have close quality to each other. Another advantage of the proposed method is the satisfactory running time of the proposed algorithm in finding a new sample. In fact, perhaps the first question for asking is about time complexity and relatively slow convergence of the bee colony algorithm. In response, due to the conscious selection of important nodes and using them in the initial solutions, it generates high quality solutions for the bee colony algorithm in terms of fitness function calculation. The experimental results on real world networks show that the proposed method is the best to preserve the degree distribution parameters, clustering coefficient, and community structure in comparison to other method.**

**Keywords**: **Sampling, Social networks, Clustering coefficient, Artificial Bee Colony**

## ۱- مقدمه

امروزه با پیشرفت فناوری، الگوهای تعاملی افراد با یکدیگر به‌طور بارزی تغییر یافته است. شبکه‌های اجتماعی مجازی و برخط یکی از ابزارهای تعاملی اجتماعی است که در همین اواخر بسیار همه‌گیر و موجب برقراری ارتباط با مسافت طولانی و با هزینه کم شده است [1]. گستردگی استفاده از شبکه‌های اجتماعی مانند فیس‌بوک و توییتر باعث شده که توجه پژوهش‌گران از جنبه‌های مختلفی به تجزیه و تحلیل این شبکه‌های اجتماعی حجیم معطوف شود [2, 3]. شبکه‌های اجتماعی، با میلیون‌ها کاربر که به‌طوردائمی پویا هست، هنگام تجزیه و تحلیل نیازمند صرف هزینه زیادی از لحاظ فضای ذخیره‌سازی گره‌ها و پیچیدگی زمانی خواهد بود [2]، از طرف دیگر محدویت دسترسی به اطلاعات کاربران در شبکه‌های اجتماعی یکی دیگر از چالش‌های پیش روی تجزیه و تحلیل این شبکه‌ها است [4, 5]؛ لذا برای غلبه بر این مسائل روش نمونه‌گیری مطرح می‌شود تا یک نمونه کوچک از شبکه بسیار بزرگ تولیدشده، تجزیه و تحلیل با سرعت بالا و هزینه کم بر روی این نمونه که معرف شبکه اجتماعی اصلی است، انجام گیرد [6]. نمونه‌گیری به‌عبارتی روشی است که یک زیرمجموعه‌ای از گره‌ها یا یال‌های موجود از گراف شبکه اجتماعی مورد نظر انتخاب می‌شود [7].

الگوریتم‌های زیادی برای نمونه‌گیری از شبکه‌های اجتماعی وجود دارد؛ کارایی این الگوریتم‌ها به‌طورکلی به‌وسیله اندازه‌گیری خواص اصلی شبکه مانند توزیع درجه[1]، ضریب خوشه‌بندی[2] و غیره در نمونه شبکه به‌دست‌آمده، مورد ارزیابی قرار می‌گیرد. بررسی این خواص از نمونه شبکه تولیدشده، جهت اطمینان از حفظ خواص شبکه اصلی توسط آن الگوریتم نمونه‌گیری است [7-9]. از روش‌های محبوب و سریع برای نمونه‌گیری شبکه می‌توان به نمونه‌گیری تصادفی ساده و نمونه‌گیری با قدم‌زدن تصادفی اشاره کرد. با این حال، نشان داده شده است که برای یک شبکه اصلی با ویژگی‌هایی همچون مقیاس آزادبودن، شبکهٔ نمونهٔ حاصل از روش‌های تصادفی در بیش‌تر مواقع، از ویژگی‌های شبکه‌های مقیاس آزاد پیروی نمی‌کنند؛ حتی اگر برخی از ویژگی‌های ساختاری شبکه‌ها مثل توزیع درجه هم تا حدی مشابه با شبکه‌های اصلی باشد، باز تفاوت پارامترهای دیگر زیاد بوده و با افزایش اندازه نمونه و اندازه شبکه اصلی این تفاوت‌ها بیشتر نیز افزایش می‌یابد [10]. چالش پیش روی نمونه‌گیری این است که نمونه گراف ایجادشده از گراف اصلی شبکه باید مناسب و قابل اعتماد باشد؛ به‌طوری‌که بتواند خواص اصلی گراف شبکه را منعکس کند [11]. الگوریتم‌های نمونه‌گیری می‌توانند هم قطعی و هم احتمالی باشند که بر روی مجموعه کوچکی از پارامترها وابسته هستند و ارزیابی اعتبار آنها راهی برای تخمین چنین پارامترهایی از داده‌های مشاهده شده است [12]. تعیین اهمیت گره‌ها بر اساس ویژگی‌های ساختاری و اهمیت‌دادن به گره‌های مهم در فرایند نمونه‌برداری و نیز سعی در به‌دست‌آوردن نمونه‌هایی که بیشترین مشابهت را با شبکه اصلی داشته باشد، از چالش‌های اساسی روش‌های اخیر در نمونه‌برداری است. در این مقاله به بررسی روش‌های نمونه‌گیری از شبکه‌های

[1] Distribution Degree

[2] Clustering Coefficient





[ DOI: 10.29252/jsdp.17.3.55 ]

اجتماعی پیچیده پرداخته و روش جدیدی را برای رفع مشکل موجود در روش‌های نمونه‌گیری مبتنی بر الگوریتم‌های تکاملی مثل الگوریتم ژنتیک [13] ارائه داده‌ایم. هدف اصلی این مقاله استفاده از روش فراابتکاری کلونی زنبورعسل با رویکرد انتخاب آگاهانه گره‌ها جهت حضور در نمونه گراف تولید شده است تا دقت نمونه‌گیری افزایش یابد؛ که در روش پیشنهادی پس از تشخیص گره‌های مهم شبکه اجتماعی همانند گره هسته و هاب‌ها، سعی می‌شود تا گره‌های با اهمیت به‌حتم در نمونه گراف تولیدشده حضور داشته باشند تا نمونه حاصل از لحاظ کیفیت به گراف شبکه اصلی مورد نظر به‌طورتقریبی مشابهت داشته باشد. در اصل، با تعیین اهمیت گره‌ها در روش پیشنهادی و بسط نمونه‌برداری از گره‌های مهم و نیز درنظرگرفتن اهمیت گره‌ها در احتمال انتخاب‌شان در نمونه‌برداری باعث می‌شود نمونه حاصل، خواص اصلی شبکه‌های پیچیده مثل مقیاس آزادبودن، ضریب خوشه‌بندی پایین و اسپارس‌بودن را در گراف نمونه حاصل نیز حفظ کند که این از مزیت‌های اساسی روش پیشنهادی محسوب می‌شود.

در ادامهٔ این مقاله، در بخش دوم، پیشینه پژوهش را بیان می‌کنیم. در بخش سوم جزئیات الگوریتم و روش پیشنهادی و نحوه پیاده‌سازی آن به‌طور کامل توضیح داده خواهد شد؛ بخش چهارم شامل معرفی مجموعه‌داده‌های استاندارد مورد استفاده جهت انجام آزمایش‌ها و مقایسه نتایج نشان داده و درنهایت در بخش پنجم جمع‌بندی و ارائه نتایج کلی و پیشنهادهایی برای کارهای آینده بیان می‌شود.

## ۲- پیشینهٔ پژوهش

روش‌های زیادی برای نمونه‌گیری از گراف شبکه‌های اجتماعی ارائه شده که هدف آنها ایجاد یک نمونه گرافی است که از لحاظ خواصی مانند توزیع درجه، ضریب خوشه‌بندی، تراکم داخلی و غیره به‌طورتقریبی متناظر با گراف شبکه اصل باشد. یکی از قدیمی‌ترین روش‌های نمونه‌گیری، روش قدم‌زدن تصادفی تندرویِ حجیم نامیده می‌شود که توسط *متروپولیس* و همکارانش در سال ۱۹۵۳ ارائه شده است. این روش به‌صورت خیلی وسیع در توزیع مونت کارلو با یافتن توزیع مدنظر از یک گراف شبکه غیرجهت‌دار متصل به‌کار می‌رود و جهت گذر از یک گره به گره دیگر از احتمالاتی براساس توزیع یافته‌شده استفاده می‌کند [14]. ایراد اساسی این روش، عدم توجه به پیچیدگی زمانی بسیار زیادی است که در ساخت گراف نمونه حاصل می‌شود. روش نمونه‌گیری برفی توسط *گودمن* در سال ۱۹۶۱ ارائه شده است که این روش مشابه روش جستجوی خطی نمونه‌گیری انجام می‌داد [15]؛ همچنین *لووز* در سال ۱۹۹۳، روش نمونه‌گیری دیگری را مبتنی بر قدم‌زن تصادفی ارائه کرد [16]. در این روش ابتدا یک گره به‌صورت تصادفی به‌عنوان گره مرکزی انتخاب و سپس حول گره مرکزی گره‌های همسایه با پیاده‌روی و قدم‌زدن انتخاب و یک نمونه شبکه ایجاد می‌شود. روش نمونه‌گیری واکنش‌محور توسط *هکسورن* در سال ۱۹۹۷ ارائه شد [17]. این روش نمونه‌گیری به‌عنوان یک رویکرد جدید جهت مطالعه جمعیت‌های مخفی در شبکه‌های اجتماعی مورد استفاده قرار گرفته، و ایده اصلی آن تصحیح نارسایی‌های موجود در نمونه‌گیری‌های تصادفی است. *لسکوویچ* و همکارانش در سال ۲۰۰۶ روش‌های نمونه‌گیری را در دو گروه تصادفی و اکتشافی طبقه‌بندی کردند [9]. روش‌های تصادفی خود به دو صورتِ گره تصادفی و یال تصادفی شناخته می‌شوند که به‌ترتیب با انتخاب تصادفی گره‌ها و یال‌ها نمونه گراف‌هایی را تولید می‌کنند؛ اما در این دو روش به‌دلیل ماهیت انتخاب تصادفی، نمونه گراف‌های تولیدشده ویژگی‌های مورد نظر گراف اصلی را منعکس نمی‌کنند. همین‌طور روش نمونه‌گیری مبتنی بر رتبه‌بندی گره نیز توسط ایشان ارائه شد که نمونه‌گیری و انتخاب گره‌ها براساس رتبه و امتیاز گره‌ها صورت می‌گیرد و احتمال انتخاب‌شدن گره‌هایی با رتبه بالاتر بیشتر از سایر گره‌ها است که این روش منجر به تولید گراف‌های متراکم‌تر از گراف اصلی و ناسازگار با گراف اصلی است. روش دیگری که توسط *لسکوویچ* و همکارانش ارائه شده، روش نمونه‌گیری براساس درجه گره است که بیشتر در نمونه‌گیری گره‌های با درجه بالاتر انتخاب می‌شوند؛ بنابراین بازنمونهٔ گراف تولیدشده متراکم‌تر از گراف اصلی و ناسازگار با آن خواهد بود و نمی‌تواند ویژگی‌های گراف اصلی را منعکس کند [9]. روش نمونه‌گیری مبتنی بر خزیدن[1] توسط *شائوژی* و همکارانش در سال ۲۰۱۰ ارائه شد [18]. در این روش ابتدا گره مرکزی به‌صورت تصادفی انتخاب و سپس این گره به صف انتخاب گره منتقل می‌شود، گره v از صف، انتخاب و به فهرست گره‌های انتخاب‌شده در نمونه‌گیری انتقال داده می‌شود؛ سپس انتقال گره‌های جدید به صف و تکرار مراحل اضافه‌کردن گره تا رسیدن به تعداد گره‌های مورد نظر در نمونه گراف ادامه خواهد یافت. چهار

[1] Crawling





[ DOI: 10.29252/jsdp.17.3.55 ]

الگوریتمی که برای انتخاب گره و خزش آن مورد استفاده قرار می‌گیرد، عبارتند از: جستجوی سطحی، عمقی، انتخاب گره متناسب با احتمال درجه آن و یا انتخاب گره‌هایی با درجهٔ پایین است [18]. روش نمونه‌گیری مبتنی بر کوتاه‌ترین مسیر توسط *رضوانیان* و *میبدی* در سال ۲۰۱۴ ارائه شده است که در این روش با استفاده از الگوریتم دیکسترا کوتاه‌ترین مسیر بین دو گره انتخابی پیدا شده، و براساس تعداد تکرار ظهور هر یال به آنها امتیازی داده و نمونه با انتخاب یک درصد معینی از گره‌ها با رتبه بالا ایجاد می‌شود [19]. روش نمونه‌گیری مبتنی بر الگوریتم ژنتیک از شبکه و داده مقیاس آزاد توسط *پاول کرومر و جان پلوتس* ارائه شده است، تا از گراف شبکه‌های اجتماعی به روش بهینه‌سازی‌شده و با درنظرگرفتن توزیع قانون توانی[1] نمونه‌گیری انجام گیرد [13]. از نتایج به‌دست‌آمده از این روش کارایی بالای آن نسبت به الگوریتم‌های تصادفی از لحاظ حفظ توزیع درجه در نمونه به‌دست آمده را می‌توان اشاره کرد؛ اما انتخاب جمعیت اولیه و نمونه‌های اولیه به‌صورت تصادفی از معایب این روش است. این روش کارایی بالاتری نسبت به روش‌های نمونه‌گیری تصادفی از لحاظ خطای نسبی آزمایش *کولموگرف - اسمیرنوف* دارد. از معایب این روش می‌توان به پیچیدگی زمانی بالای آن نسبت به سایر روش‌های نمونه‌گیری تصادفی اشاره کرد. روش نمونه‌گیری سانتریفوژ قدم‌زن تصادفی توسط *سویلا* و همکارانش در سال ۲۰۱۵ ارائه شده است [20]. این روش برای نمونه‌گیری از شبکه‌هایی با توزیع احتمال ثابت مورد استفاده قرار گرفت. در این روش با انتخاب تصادفی گره‌ای به‌عنوان گره مرکزی و قدم‌زدن در حول گره مرکزی یک درخت پوشا ایجاد کرده که در این درخت پوشا وزن هر گره به‌عنوان معیاری جهت انتخاب آن در نمونه‌گیری مورد استفاده قرار می‌گیرد. از مزایای این روش می‌توان به تولید نمونه گرافی با توزیع احتمال مشابه گراف اصلی با پیچیدگی زمانی پایین اشاره کرد. روش نمونه‌گیری آتش‌سوزی جنگل بر اساس رتبه‌بندی صفحه توسط *تانگ* و همکارانش در سال ۲۰۱۵ مطرح شد [21]. در این روش نمونه‌گیری بر روی حفاظت ساختار اجتماعات کوچک و همچنین حفظ ساختار شبکه اصلی متمرکز می‌شود؛ چون ساختار انجمن، توزیع نابرابر گره‌ها و تجاوز یال‌ها را نمایش می‌دهد. روش نمونه‌گیری آتش‌سوزی جنگل را با دو گام اساسی بهبود بخشیده است که در نخستین گام بعد از پارتیشن‌بندی گراف به انجمن‌هایی، برای هر گره یک ضریب اجتماع محاسبه و مرکز خوشه انتخاب و در یک مجموعه‌ای به نام مراکز خوشه‌ها جمع‌آوری می‌شود. در دومین گام، با استفاده از الگوریتم رتبه‌بندی صفحه، برای هر گره مقدار رتبه (وزن) محاسبه می‌شود؛ بعد در مراحل بعدی با توجه به گره‌ای که بزرگ‌ترین ضریب خوشه‌بندی را از مجموعه مراکز دارد، از خوشه مربوطه آن نمونه‌گیری را از گره با کمترین رتبه شروع می‌کند. از مزایای این روش می‌توان به حفظ نسبت گرهٔ یال، حفظ تقریبی ساختار انجمنی و ارتباطات اشاره کرد [21].

الگوریتم نمونه‌برداری TIES به‌کرار یک یال را به‌صورت تصادفی از گراف شبکهٔ اجتماعی انتخاب کرده و گره‌های مبدأ و هدف مربوط به یال جاری را در هر تکرار به گراف نمونه اضافه می‌کند [22]. در گام بعدی، الگوریتم با جستجو در گراف اصلی، تمامی یال‌های ممکن مابین گره‌های انتخاب‌شده از گراف اصلی را به گراف نمونه اضافه می‌کند و درنهایت الگوریتم زمانی متوقف می‌شود که کسری معین از گره‌های تعیین‌شده از گراف شبکهٔ اجتماعی به گراف نمونه اضافه شده باشند. بردلر و همکارانش یک روش برای برآورد تعداد مطابقت کامل گراف‌های غیر جهت‌دار براساس رابطه بین نمونه‌گیری بوسون گاوسی[2] و تئوری گراف پیشنهاد کردند [23]. آنها کدگذاری‌هایی را از ماتریس مجاورت یک گراف به یک حالت گاوسی ارائه داده و استراتژی‌هایی را برای افزایش احتمال موفقیت در نمونه‌برداری برای گراف‌های مورد مطالعه نشان دادند. در یک کار دیگری، روش نمونه‌برداری بر اساس گراف‌های هدف بیان می‌شود. به‌عنوان مثال، ژائو و همکارانش روش‌هایی را توسعه داده‌اند که می‌توانند بدون نیاز به نمونه‌برداری یک‌نواخت رئوس، یک گراف را به‌طور مؤثر نمونه‌برداری کنند [24]. در بسیاری از مواقع، یک گراف هدف با یک گراف کمکی و یک گراف دوبخشی مرتبط است و آنها در کنار هم ساختار شبکه متصل دولایهٔ بهتری را تشکیل می‌دهند. این دیدگاه جدید مزایای اضافی را برای نمونه‌برداری از گراف به‌همراه دارد. اگر نمونه‌برداری مستقیم از یک گراف هدف دشوار باشد، می‌توان آن را به‌طور غیرمستقیم با کمک دو گراف دیگر نمونه‌برداری کرد؛ همچنین دو استراتژی جدید مبتنی بر قدم‌زدن تصادفی ارائه شده است که به‌ترتیب CSN و NR نامیده شده‌اند [25]. استراتژی‌های CSN و NR به‌ترتیب بر روی تأثیر انتخاب گرهٔ دانه و هم‌پوشانی مسیر تمرکز می‌کنند. این دو استراتژی نسبت به دیگر روش‌های

[1] Power Law

[2] Gaussian boson





[ DOI: 10.29252/jsdp.17.3.55 ]

مبتنی بر قدم‌زدن تصادفی، کارایی به‌نسبه بهتری در برخی شبکه‌های مصنوعی مثل BA[1] و WS[2] دارد. در این مقاله ما روش خود را با هر دوی اینها مقایسه کرده‌ایم.

## ۳- روش پیشنهادی

در این مقاله برای نخستین‌بار الگوریتم کلونی زنبورعسل به‌همراه یک رابطه جدید میتنی بر ویژگی‌های ساختاری شبکه‌های اجتماعی جهت نمونه‌گیری از شبکه‌های اجتماعی پیاده‌سازی شده است. با توجه به ماهیت تصادفی‌بودن انتخاب جمعیت اولیه در روش‌های نمونه‌گیری فراابتکاری مثل الگوریتم ژنتیک و دیگر الگوریتم‌های تکاملی، ما در روش پیشنهادی این مقاله، ایده انتخاب آگاهانه گره‌ها جهت تولید نمونه‌های اولیه (راه‌حل‌های اولیه) با استفاده از روش یافتن گره‌های هاب و شبه‌هاب و نیز گرهای مهم دیگر را مثل هسته‌ها به‌منظور تلاش برای حضور گره‌های مهم در نمونه به‌دست آمده، ارائه داده‌ایم، تا درنهایت نمونه باکیفیت بالا و نزدیک به کیفیت شبکه اصلی از لحاظ ویژگی‌ها به‌دست آید. در این روش نمونه گراف‌های به‌دست‌امده با شبکه اصلی سازگاری خوبی داشته و می‌تواند خصوصیات شبکه اصلی از جمله توپولوژی و حتی تعداد اجتماعات و هم‌بندبودن مؤلفه عظیم گراف نمونه اصلی را تا حد ممکن حفظ می‌کند. در جدول (۱) اصطلاحات معادل الگوریتم کلونی زنبور عسل با نمونه‌گیری از شبکه‌های اجتماعی، نمایش داده شده و در ادامه الگوریتم پیشنهادی گام به گام تفسیر شده است.

**(جدول-۱): اصطلاحات کلونی زنبور عسل در نمونه‌گیری**
**(Table-1): Terms of bee colony in sampling**

| نمونه‌گیری | کلونی زنبورعسل |
|---|---|
| نمونه گراف | Bee |
| تعداد کل نمونه‌ها | NB |
| تعداد نمونه‌های اولیه | NF |
| نشان‌دهنده شماره برچسب نخستین گره گراف (معمولا از ۱ شروع می‌شود.) | Lower bound |
| نشان‌دهنده شماره برچسب آخرین گره گراف | Upper bound |
| تعداد گذرها | Iteration |
| محدودیت | Trial |
| میزان کیفیت | Cost-Fit |

مراحل الگوریتم پیشنهادی مبتنی بر کلونی زنبور عسل به‌صورت گام به گام در زیر بیان می‌شود:

**ورودی**: گراف‌های مربوط به شبکه‌های اجتماعی

**خروجی**: نمونه گراف کوچک تولیدشده از روی گراف اصلی

۱- تنظیمات اولیه را انجام بده:

a. ایجاد ماتریس همسایگی از گراف شبکه و تعیین اندازه شبکه اصلی؛

b. تعیین اندازه نمونه‌ها که در این الگوریتم درصدی از کل گره‌های شبکه را در نظر می‌گیریم؛

c. تعیین تعداد نمونه شبکه‌هایی که تولید می‌شود؛

d. تعیین برچسب گره پایین و برچسب گره بالا و تعداد تکرار الگوریتم.

۲- تعداد NB زنبور (نمونه) ایجاد کن که در ابتدا هر زنبور عسل یک آرایه k عضوی است (هر کدام از این نمونه‌ها دارای k تا گره خواهند بود).

۳- تعداد NF زنبور را به‌دنبال منابع غذایی ارسال کن (به‌طورمعمول NF را نصف NB در نظر می‌گیریم).

۴- مقدار حلقه را برای L گام تنظیم کن (به‌طورمعمول L=100)

۵- مقدار ضریب تنبیه[3] که زنبور عسل به‌اندازهٔ محدودی اطراف یک منبع غذایی حضور داشته باشد و درصورت عدم یافتن منبع غذایی بعد از تعداد تکرار معینی بهتر است، آن محل را ترک کند و در محل دیگری حضور یابد.

۶- ابتدا برای گراف اولیه یک مقدار آلفا و Xmin براساس درجه گره‌ها محاسبه کن.

۷- فراخوانی تابع محاسبه مقادیر هسته[4] و هاب[5] برای تمامی گره‌ها جهت انتخاب آگاهانه گره‌ها در فرایند نمونه‌گیری.

$$\text{Corenode(i)}=\frac{\text{Di}+\sum_{j=1}^{m} dij}{M} \quad (۱)$$

۸- فرمول (۱) تا حد بسیار زیادی به‌دست‌آوردن ارزش هسته‌بودن گره است که $D_i$ مقدار درجه گره i بوده، $\sum_{j=1}^{m} dij$ مجموع درجه همسایه‌های گره i و M نیز تعداد همسایه‌های گره i است.

$$\text{Hubnode(i)}=D_i*\frac{1}{CCi+0.01}+\frac{\sum_{i}^{m} dji}{m} \quad (۲)$$

[1] Barabási-Albert
[2] Watts-Strogtz

[3] Trial
[4] Core
[5] Hub




[ DOI: 10.29252/jsdp.17.3.55 ]

۹- در فرمول (۲)، مقدار $Hubliknode(i)$ نشان‌دهنده میزان شباهت گره به گره‌های hub در شبکه است، $d_i$ درجه گره $i$ام، $\sum_{j=1}^{M} d_{ij}$ مجموع درجه گره‌های همسایگان گره $i$، M تعداد همسایگان گره $i$، و $CC_i$ ضریب خوشه‌بندی هر گره است. با توجه به اینکه ضریب خوشه‌بندی می‌تواند صفر باشد، بنابراین جهت جلوگیری از صفربودن مخرج یک عدد ثابت ۰/۰۱ به آن اضافه کرده‌ایم.

۱۰- نصف نمونه‌ها (زنبورهای کارگر) به‌صورت تصادفی و همانند روش پایه الگوریتم کلونی زنبور عسل تولید می‌شوند.

۱۱- نصف دیگر نمونه‌ها به‌طورکامل آگاهانه تولید می‌شوند. گره‌هایی در این نمونه‌ها حضور خواهند یافت که مقدار هسته یا هاب‌بودن بیشتری داشته باشند؛ البته تمامی گره‌ها با احتمال مساوی، شانس انتخاب‌شدن را دارند؛ اما براساس چرخ رولت سعی کردیم گره‌هایی که مقدار هاب یا هسته بیشتری دارند، شانس بیشتری نیز در انتخاب‌شدن داشته باشند، و گره‌های با مقدار کمتر نیز احتمال انتخاب کمتری دارند. این استراتژی برای افزایش احتمال انتخاب گره‌های مهم و حضور آنها در نمونه شبکه به‌دست آمده است تا از لحاظ توزیع درجه و حفظ تقریبی ساختارهای متراکم با حفظ گره‌های مهم و اساسی، شباهت زیادی به شبکه اصلی داشته باشد.

۱۲- در این مرحله مقدار Fit نمونه‌های تولیدشده از طریق تابع سودمندی[1] محاسبه می‌شود. هدف از این مرحله بررسی نمونه تولیدشده از لحاظ توزیع قانون توانی بوده تا نمونه‌ای که کمترین مقدار Fit را داشته باشد، به‌عنوان نمونه‌ای که شباهت بیشتری به گراف اصلی دارد، مورد قبول باشد.

۱۳- در هر مرحله کمترین مقدار Fit مربوط به نمونه، همراه با ایندکس آن، ذخیره می‌شود و نمونه‌ای بهترین نمونه خواهد بود که از نظر مقدار آلفا و Xmin کمترین اختلاف را با مقدار آلفا و Xmin گراف اصلی داشته باشد.

یکی از تفاوت‌های بارز شبکه‌های پیچیده، داشتن گره‌های با ویژگی‌های توپولوژیکی خاص مثل گره‌های هسته و هاب با تعداد بسیار کم است که اساس ساختاری شبکه‌های پیچیده‌ای مثل شبکه‌های اجتماعی وابسته به این گره‌ها هستند. به‌عبارتی، اگر این گره‌های مهم را از شبکه حذف کنیم، شبکهٔ مورد نظر، دیگر شبکه پیچیده نخواهد بود. به همین دلیل ما با درنظرگرفتن رابطه (۱ و ۲) سعی می‌کنیم تا ساختار شبکه پیچیده را تا حد ممکن در شبکه نمونه‌برداری‌شده جدید حفظ کنیم. این کار به‌صورت مستقیم باعث حفظ ویژگی مقیاسِ آزادبودن شبکه می‌شود و به‌صورت غیرمستقیم نیز وجود ساختارهای جوامع را در گراف حفظ می‌کند. رابطه دوم روی اهمیت‌دادن به گره‌های هاب و پل تأکید دارد. در صورت نادیده‌گرفتن اهمیت چنین گره‌هایی، گراف نمونه حاصل، به مؤلفه‌های مختلفی شکسته می‌شود و بنابر این نمونه حاصل، نماینده خوبی برای شبکه اصلی نخواهد بود.

نمونه‌های اولیه با انتخاب تصادفی گره‌ها براساس چرخ رولتی که برمبنای مقادیر به‌دست‌آمده از فرمول‌های بالا تشکیل شده، به‌وجود می‌آیند. در این مرحله با مرتب‌سازی و کنترل گره‌های انتخاب‌شده، از تولید گره تکراری در هر نمونه جلوگیری می‌شود. مقادیر متغیرهای $\alpha^0$ و $X_{min}^0$ برای شبکه اصلی بعد از به‌دست‌آوردن درجه گره‌های شبکه اصلی و ارسال آن به‌عنوان پارامتر به تابع توزیع قانون توانی که پیاده‌سازی شده است، به‌دست می‌آید. مقادیر $\alpha^c$ و $X_{min}^c$ نیز برای نمونه شبکه‌های تولید شده است که بعد از به‌دست‌آوردن درجه گره‌های نمونه شبکه مورد نظر و ارسال آن به‌عنوان پارامتر به تابع توزیع قانون توانی، به‌دست می‌آیند. مقدار کیفیت هر نمونه توسط تابع سودمندی، با دریافت مقادیر $\alpha^0$ و $X_{min}^0$ از شبکه اصلی و مقادیر $\alpha^c$ و $X_{min}^c$ از نمونه شبکه مورد نظر به‌عنوان ورودی طبق فرمول (۵.۳) محاسبه می‌شود. در این مرحله برای هر نمونه شبکه یک مقدار کیفیت محاسبه می‌شود.

$$Fit(c)=\frac{1}{2}\left(\sqrt{(\alpha^0-\alpha^c)^2}+\sqrt{(X_{min}^0-X_{min}^c)^2}\right) \quad (۳)$$

تا این مرحله نمونه‌شبکه‌های اولیه‌ای تولید شده‌اند، که شامل یک مقدار کیفیتی نیز هستند. در این مرحله ابتدا دو نمونه‌شبکه به نام‌های $i$ و p به‌صورت تصادفی از نمونه‌شبکه‌های تولیدشده در مراحل قبلی، انتخاب و سپس در هر دو نمونه‌شبکهٔ بعدی از آنها برای مثال گره jام در بردار k عضوی زنبور نمونه به تصادف انتخاب و بر طبق فرمول زیر یک برچسب جدید برای آن نمونه محاسبه می‌شود (حرکت به سمت گره جدید و تولید نمونه جدید).

$$Newfood.par(j)= (newfood.par(j)+unifrnd(-1,1)* (newfood.par(j)-food(p).par(j))) \quad (۴)$$

[1] Fitness Function





[ DOI: 10.29252/jsdp.17.3.55 ]

بعد از تولید نمونه‌شبکه جدید، مقدار کیفیت نمونه‌شبکه جدید محاسبه می‌شود؛ اگر کیفیت نمونه‌شبکه جدید از نمونه‌شبکه $i$ بهتر باشد، نمونه‌شبکه جدید جایگزین آن نمونه‌شبکه می‌شود، وگرنه نمونه‌شبکه دیگری تولید خواهد شد و به مقدار Trial نمونه‌شبکه $i$ یک واحد اضافه می‌شود؛ و این روند تا رسیدن به تعداد مورد نظری از نمونه-شبکه‌ها، ادامه می‌یابد. در گام بعدی جهت شبیه‌سازی حرکت زنبورهای ناظر با تولید نمونه‌شبکه‌های جدید براساس میزان کیفیت نمونه‌شبکه‌های قبلی صورت می‌گیرد. ابتدا چرخ رولتی براساس میزان کیفیت نمونه‌شبکه‌های موجود ایجاد کرده، تا نمونه‌شبکه‌های باکیفیت شانس بالاتری برای انتخاب‌شدن داشته باشند؛ سپس براساس این چرخ رولت دو نمونه شبکه‌جدید به نام‌های $i$و p انتخاب می‌کنیم، در این مرحله باز نمونه‌شبکهٔ جدیدی براساس فرمول (۴) ایجاد می‌کنیم. بعد از تولید، از لحاظ میزان کیفیت با نمونه‌شبکه $i$ مورد مقایسه قرار می‌گیرد؛ اگر کیفیت نمونه‌شبکهٔ جدید از نمونه‌شبکه $i$ بهتر باشد، نمونه‌شبکهٔ جدید جایگزین آن نمونه‌شبکه می‌شود، وگرنه نمونه‌شبکه دیگری تولید خواهد شد و به مقدار Trial نمونه‌شبکه $i$ یک واحد اضافه می‌شود. و این روند تا رسیدن به تعداد مورد نظری از نمونه‌شبکه‌ها، ادامه می‌یابد. در ادامه، نمونه‌شبکه‌هایی را که مقدارTrial آنها از محدوده تعیین‌شده بیشتر است، شناسایی و سپس نمونه‌های جدیدی برطبق مرحله تولید نمونه شبکه‌های در مرحلهٔ نخست، ایجاد می‌کنیم تا جایگزین آنها شوند (شبیه‌سازی زنبورهای پیشرو). در هر بار اجرا بهترین نمونه‌شبکه از لحاظ کیفیت، حفظ می‌شود تا زمانی‌که به انتهای حلقه رسیدیم. درنهایت، نمونه‌شبکه‌ای که بهترین کیفیت یا به‌عبارتی کمترین اختلاف را از نظر توزیع درجه با شبکه اصلی داشته باشد، تولید می‌شود.

یک مزیت مهم روش پیشنهادی، پایداری روش جدید در استخراج نمونه‌های باکیفیت در هر بار اجرا است. به‌عبارتی، برخلاف رفتار تصادفی الگوریتم زنبور عسل، همواره نمونه‌های حاصل کیفیت بسیارنزدیکی به هم دارند و از نظر ویژگی‌های محاسبه‌شده که در بخش نتایج اشاره شده‌اند، به‌طورکامل به هم نزدیک هستند. دلیل این کار هم انتخاب آگاهانه و غیرتصادفی در گام‌های اولیه بسط و توسعه نمونه‌ها است.

یکی دیگر از مزایای روش پیشنهادی، سرعت خوب الگوریتم پیشنهادی در یافتن شبکه نمونه‌ای جدید است. درواقع، شاید نخستین سؤالی که به ذهن برسد درخصوص پیچیدگی زمانی و هم‌گرایی به‌نسبه کُند الگوریتم کلونی زنبور عسل باشد؛ درحالی‌که روش ما به‌دلیل انتخاب آگاهانه گره‌ها، جواب‌های اولیه‌ای که برای الگوریتم کلونی زنبور عسل تولید می‌کند، کیفیت بالایی از نظر محاسبه تابع سودمندی دارند. آزمایش‌ها نشان داد که الگوریتم کلونی زنبور عسل در کمتر از ده تکرار، به جواب نهایی هم‌گرا می‌شود و بنابراین طبق گام توسعه نمونه بر اساس گره‌های مهم و همسایگی آنها، بیشینه زمان پیچیدگی برابر است با $O(Ik^2n)$، که $I$ تعداد تکرار را نشان می‌دهد و $k$ متوسط درجات گره‌ها را بیان می‌کند. به‌دلیل مقدار بسیار ناچیز $d$ و با درنظرگرفتن ویژگی $kn=m$ در شبکه‌های واقعی، پیچیدگی روش ارائه‌شده، جزء روش‌های نمونه‌برداری به‌طورتقریبی خطی[1] است که برابر است با $O(km)$، که m نشان‌دهندهٔ تعداد یال‌های شبکه است.

## ۴- ارزیابی روش پیشنهادی

برای اثبات کارایی الگوریتم پیشنهادی در نمونه‌گیری از شبکه‌های مقیاس آزاد مثل انواع مختلف شبکه‌های اجتماعی و پیچیده، آزمایش‌هایی روی مجموعه‌داده استاندارد شبکه‌های واقعی مختلف که در جدول (۲) نشان داده شده است، انجام شد. در جدول (۳) پارامترهای نمونه‌گیری برای دو روش تکاملی الگوریتم نمونه‌گیری ژنتیک و الگوریتم پیشنهادی مبتنی بر زنبور عسل نشان داده شده است. برای سایر روش‌های مقایسه‌شده مثل الگوریتم نمونه‌گیری تصادفی، الگوریتم TIES [22]، روش CSN و NR [25] سعی کردیم تا در شرایط به‌طورکامل یکسان الگوریتم‌ها را اجرا کنیم. در جداول بعدی مقادیر به‌دست‌آمده در طی آزمایش‌های صورت‌گرفته با نرخ نمونه‌گیری ۲۰٪و با تعداد گذر صد مرحله نشان داده می‌شود. در این مقاله الگوریتم پیشنهادی نمونه‌گیری همانند سایر روش‌ها، از لحاظ پارامترهای مهم مثل توزیع توانی، میانگین ضریب خوشه‌بندی، خطای نسبی و غیره مورد مقایسه و بررسی قرار می‌گیرد.

برای بررسی کارآیی روش پیشنهادی، مقایسهٔ بین روش پیشنهادی با الگوریتم‌های پایه‌ای که در این زمینه وجود دارد، صورت گرفت. همان‌طورکه از جداول زیر

[1] Near linear





[ DOI: 10.29252/jsdp.17.3.55 ]

پیداست، روش پیشنهادی نسبت به الگوریتم‌های پایه، دارای نتایج بهتری است، که نشان‌دهندهٔ دقت بالای روش پیشنهادی در نمونه‌گیری از گراف شبکه‌های اجتماعی است.

**(جدول-۲): شبکه‌های واقعی مورد آزمایش قرار گرفته به‌همراه Xmin و$\alpha$مقادیر پارمترهای**

**(Table-2): Real-world networks with values of α and Xmin parameters**

| ردیف | نام شبکه | تعداد گره | تعداد یال | مقدار $\alpha$ | مقدار $X_{min}$ |
|---|---|---|---|---|---|
| ۱ | Netscience | 1461 | 2742 | 3.41 | 4 |
| ۲ | Internet_Routers-22july06 | 22963 | 48436 | 2.81 | 5 |
| ۳ | Karate | 34 | 78 | 2.12 | 2 |
| ۴ | Hep-th | 8360 | 113689 | 3.45 | 10 |
| ۵ | Football | 115 | 613 | 3.43 | 8 |
| ۶ | Dolphins | 62 | 159 | 3.47 | 6 |
| ۷ | Cond-mat | 16726 | 47594 | 3.48 | 16 |
| ۸ | Astro-ph | 16704 | 121251 | 3.45 | 45 |
| ۹ | Lesmis | 77 | 254 | 3.21 | 12 |
| ۱۰ | Celegansneural | 297 | 1017 | 3.36 | 11 |
| ۱۱ | Polbooks | 105 | 411 | 2.64 | 5 |

**(جدول-۳): پارامترها و تنظیمات اولیه مورد نیاز الگوریتم‌های نمونه‌گیری**

**(Table-3): Parameter settings for testing of sampling algorithms**

| | | |
|---|---|---|
| **الگوریتم نمونه‌برداری پیشنهادی (بخش کلونی زنبور عسل)** | اندازه نمونه | ۲۰، ۳۰ و ۵۰ درصد |
| | تعداد نسل‌ها | ۱۰۰ |
| | درصد جهش | ۵۰ |
| | درصد ترکیب در نسل‌های حاصل | ۵ |
| | تکرار حلقه-گذر | ۱۰۰ |
| **الگوریتم نمونه‌برداری ژنتیک** | اندازه نمونه | ۲۰، ۳۰ و ۵۰ درصد |
| | تعداد نسل‌ها | ۱۰۰ |
| | درصد جهش | ۰/۲ |
| | درصد ترکیب در نسل‌های حاصل | ۰/۸ |
| | تکرار حلقه-گذر | ۱۰۰ |

## ۱-۴- آنالیز براساس مقادیر α و $X_{min}$

شبیه‌بودن مقادیر α و $X_{min}$ گراف نمونه به گراف اصلی شبکه، نشان‌دهنده حفظ توزیع درجه قانون توانی در نمونه گراف به‌دست است. جداول (۴ و ۵) به‌ترتیب مقادیر α و $X_{min}$ شبکه اصلی را قبل از نمونه‌برداری و بعد از نمونه‌برداری با روش پیشنهادی و الگوریتم‌های دیگر مقایسه‌شده نشان می‌دهد. نتایج به‌دست‌آمده در جدول (۴) نشان می‌دهد که مقادیر α برای روش پیشنهادی مطابقت بالایی با مقادیر α شبکه اصلی در همه شبکه‌های واقعی به‌جز مجموعه‌دادگان Dolphins و Internet_Routers-22july6 دارد. در رتبه دوم این مقایسه، روش CSN قرار دارد. بدترین نتایج به‌دست‌آمده نیز مربوط به روش نمونه‌برداری تصادفی است.

جدول (۵) نیز نشان می‌دهد که مقادیر $X_{min}$ روش پیشنهادی در مقایسه با دیگر روش‌ها مطابقت بالایی با $X_{min}$ شبکه اصلی در همه مجموعه‌دادگان به‌جز مجموعه‌دادگان Internet_Routers-22july6 دارد. در این مقایسه، روش NR در رتبه دوم و نمونه‌برداری تصادفی در آخرین رتبه قرار دارد. همان‌طورکه در جداول (۴ و ۵) مشاهده می‌کنید، نمونه گراف به‌دست‌آمده در روش نمونه‌گیری پیشنهادی ثابت می‌کند که از لحاظ ویژگی توپولوژیکی توزیع درجه قانون توانی به گراف اصلی خیلی شبیه بوده و اختلاف اندکی باهم دارند که این نشان‌دهنده حفظ توزیع درجه قانون توانی در نمونه گراف به‌دست‌آمده است؛ درحالی‌که در نمونه گراف‌های به‌دست‌آمده از طریق روش‌های دیگر مثل روش نمونه‌گیری تصادفی یا روش نمونه‌برداری با الگوریتم ژنتیک نمونه استخراج شده، با گراف اصلی باهم تا حدودی تفاوت داشته و این نشان‌دهنده عدم حفظ توزیع درجه در نمونه گراف به‌دست‌آمده است.

## ۲-۴- آنالیز براساس آزمایش کولموکروف اسمیرنوف(ks)

آنالیز براساس آزمایش کولموکروف اسمیرنوف(ks) درواقع معیار عمومی دیگری برای مقایسه توزیع دو گراف اصلی و نمونه گراف به‌دست‌آمده است. درواقع، این آزمون هم‌گونی بین فراوانی‌های تجمعی را در دو توزیع گراف اصلی و گراف نمونه‌برداری‌شده بررسی می‌کند. نتایج به‌دست‌آمده در بازه [0,1] هستند که هر چه مقدار به صفر نزدیک‌تر باشد، پس توزیع هر دو گراف اصلی و نمونه به‌دست‌آمده به هم نزدیک‌تر خواهند بود و هرچه به یک نزدیک‌تر باشد، پس توزیع دو گراف اصلی و نمونه باهم تفاوت دارند. این معیار بر اساس رابطهٔ ریاضی زیر تعریف می‌شود:

$$D = \max_x \{|F'(x) - F(x)|\} \qquad (۵)$$

در رابطهٔ بالا، x مقادیر مربوط به توزیع درجات در دو گراف مد نظر و آیتم‌های F و F' مربوط به دو تابع توزیع تجمعی درجه در گراف اصلی و نمونه اشاره دارند. جدول (۶) نشان می‌دهد که الگوریتم نمونه‌گیری پیشنهادی ما با نرخ نمونه‌گیری بیست درصد بهترین عملکرد را دارد و به





[ DOI: 10.29252/jsdp.17.3.55 ]

استثنای مجموعه‌دادگان Internet_routers-22july06، در مابقی شبکه‌ها، مقادیر کوچک‌تری (نزدیک به صفر) را نسبت به بقیه الگوریتم‌ها داشته که این نشان‌دهندهٔ اختلاف کم بین توزیع درجه گراف اصلی و نمونه گراف به‌دست‌آمده از طریق الگوریتم پیشنهادی ما است؛ یعنی توزیع درجه مورد نظر در نمونه گراف به‌دست‌آمده از این روش نسبت به روش‌های مقایسه‌شده دیگر تا حد زیادی حفظ شده است. همین آزمایش با نرخ نمونه‌گیری سی درصد در جدول (۷) تکرار شده است که نشان می‌دهد الگوریتم نمونه‌گیری پیشنهادی به استثنای مجموعه‌دادگان Internet_routers-22july06 و Karate، مقادیر کوچک‌تری را نسبت به بقیه الگوریتم‌ها در مابقی شبکه‌ها دارد.

## ۳-۴- آنالیز بر اساس حفظ ضریب خوشه‌بندی

جدول (۸) کارآیی بالا و برتری الگوریتم پیشنهادی را در حفظ ضریب خوشه‌بندی در نمونه به‌دست‌آمده از طریق نمونه‌گیری بیست درصد با روش پیشنهادی نسبت به دیگر روش‌های مقایسه‌شده در بیشتر شبکه‌ها نشان می‌دهد. همان‌طورکه مشاهده می‌شود، روش پیشنهادی در هشت مجموعه‌داده از یازده مجموعه‌داده بهترین عملکرد را دارد؛ ولی در مجموعه‌دادگان Dolphins و Cond-mat، الگوریتم CSN بهتر بوده و در Netscience نیز الگوریتم NR عملکرد بهتری در حفظ ضریب خوشه‌بندی گراف دارد. جدول (۹) نشان می‌دهد که کارآیی الگوریتم پیشنهادی در حفظ ضریب خوشه‌بندی در نمونه به‌دست‌آمده با نمونه‌گیری سی درصد نسبت به دیگر روش‌های مقایسه‌شده در هشت شبکه از یازده شبکه آزمایش‌شده بالا است. در بین روش‌های دیگر نیز الگوریتم CSN بهترین عملکرد را در دو شبکه Cond-mat و Lesmis دارد و نیز الگوریتم TIES نیز با اختلاف بسیار ناچیز ۰/۰۰۱ نسبت به روش پیشنهادی، بهترین عملکرد را در مجموعه‌داده Celegansneural دارد.

**(جدول-۴): مقادیر α به‌دست‌آمده برای روش پیشنهادی و مقایسه آن با α شبکه اصلی و الگوریتم‌های دیگر با نرخ نمونه‌گیری ۲۰٪**

**(Table-4): The obtained α values for the proposed method and compared it with α of main networks and compared algorithms at a sampling rate of 20%**

| | شبکه αمقادیر اصلی | روش پیشنهادی | CSNالگوریتم | NRالگوریتم | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Netscience** | 3.41 | **3.3438** | 2.3102 | 2.3179 | 2.7118 | 2.9118 | 1.6102 |
| **Internet_Routers-22july6** | 2.81 | **2.0777** | 2.482 | **2.0**777 | 1.6777 | **2.0**777 | 0.1518 |
| **Karate** | 2.12 | **2.1571** | 3.0082 | 3.275 | 3.0395 | 3.375 | 0.6272 |
| **Hep-th** | 3.45 | **3.444** | 3.2599 | 3.5819 | 3.2819 | 3.4819 | 1.283 |
| **Football** | 3.43 | **3.3913** | 3.0949 | 3.4913 | 3.0913 | 3.3137 | 1.349 |
| **Dolphins** | 3.47 | 3.2727 | 4.1386 | 3.9222 | **3.6108** | 3.2222 | 4.1926 |
| **Cond-mat** | 3.48 | 3.4698 | 3.159 | 3.0913 | 2.9274 | **3.4913** | 1.459 |
| **Astro-ph** | 3.45 | **3.469** | 2.8918 | 3.2924 | 3.3924 | 3.4924 | 3.4924 |
| **Lesmis** | 3.21 | **3.1429** | 2.3507 | 2.55 | 2.2459 | 2.651 | 0.5507 |
| **Celegans-neural** | 3.36 | **3.3904** | 2.6684 | 2.3738 | 2.1642 | 2.5738 | 0.2684 |
| **Polbooks** | 2.64 | **2.6011** | 3.1803 | 3.3161 | 1.7161 | 1.9161 | 3.255 |

**(جدول-۵): مقادیر $X_{min}$ حاصله برای روش پیشنهادی و مقایسه آن با $X_{min}$ شبکه اصلی و الگوریتم‌های دیگر با نرخ نمونه‌گیری ۲۰٪**

**(Table-5): The obtained $X_{min}$ values for the proposed method and compared it with $X_{min}$ of main networks and compared algorithms at a sampling rate of 20%**

| | $X_{min}$مقادیر شبکه اصلی | روش پیشنهادی | CSNالگوریتم | NRالگوریتم | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Netscience** | 4 | **4** | 3 | 3 | 3 | **4** | 2 |
| **Internet_Routers-22july6** | **5** | 4 | **5** | 4 | **5** | **5** | 3 |
| **Karate** | 2 | **2** | 1 | **2** | 1 | 3 | 5 |
| **Hep-th** | 10 | **10** | 9 | 10 | 11 | 11 | 5 |
| **Football** | 8 | **8** | 7 | 7 | **8** | **8** | 3 |
| **Dolphins** | 6 | **6** | 5 | **6** | 7 | 5 | 4 |
| **Cond-mat** | 16 | **15** | 13 | 13 | 18 | 14 | 24 |



[ DOI: 10.29252/jsdp.17.3.55 ]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Astro-ph** | 45 | **44** | 42 | 43 | 42 | 41 | 30 |
| **Lesmis** | 12 | **12** | 13 | 10 | **12** | 6 | 8 |
| **Celegansneural** | 11 | **11** | 10 | **11** | 12 | 7 | 9 |
| **Polbooks** | 5 | **5** | 6 | 8 | 6 | 4 | 8 |

(جدول-۶): تحلیل توزیع درجه بر حسب مقدار KS (با نرخ بیست درصد)

**(Table-6): The analysis of degree distribution by KS value (with rate of 20%)**

| نام شبکه | الگوریتم پیشنهادی | الگوریتمCSN | الگوریتمNR | روشTIES | نمونه‌برداری با الگوریتم ژنتیک | الگوریتم نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|
| Netscience | **0.34** | 0.35 | 0.36 | 0.45 | 0.39 | 0.37 |
| internet_routers-22july06 | 0.29 | **0.28** | 0.31 | 0.33 | 0.36 | 0.57 |
| Karate | **0.26** | 0.54 | 0.45 | 0.49 | 0.44 | 0.62 |
| hep-th | **0.33** | 0.37 | 0.39 | 0.41 | 0.36 | 0.64 |
| Football | **0.40** | 0.42 | 0.41 | 0.44 | 0.48 | 0.47 |
| Dolphins | **0.23** | 0.28 | 0.27 | 0.39 | 0.34 | 0.53 |
| cond-mat | **0.26** | 0.29 | 0.32 | 0.43 | 0.56 | 0.49 |
| astro-ph | **0.30** | **0.30** | 0.31 | 0.47 | 0.32 | 0.37 |
| Lesmis | **0.23** | 0.47 | 0.51 | 0.41 | 0.57 | 0.56 |
| Celegansncural | **0.25** | 0.35 | 0.38 | 0.39 | 0.64 | 0.54 |
| Polbooks | **0.44** | 0.49 | 0.42 | 0.39 | 0.58 | 0.43 |

(جدول-۷): تحلیل توزیع درجه بر حسب مقدار KS (با نرخ سی درصد)

**(Table-7): The analysis of degree distribution by KS value (with rate of 30%)**

| نام شبکه | الگوریتم پیشنهادی | الگوریتمCSN | الگوریتمNR | روشTIES | نمونه‌برداری با الگوریتم ژنتیک | الگوریتم نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|
| Netscience | **0.27** | 0.35 | 0.35 | 0.32 | 0.32 | 0.37 |
| internet_routers-22july06 | 0.26 | **0.25** | 0.29 | 0.41 | 0.27 | 0.36 |
| Karate | 0.25 | 0.54 | 0.44 | 0.37 | **0.15** | 0.18 |
| hep-th | **0.33** | 0.35 | 0.38 | 0.57 | 0.58 | 0.47 |
| Football | **0.26** | 0.34 | 0.30 | 0.38 | **0.26** | 0.36 |
| Dolphins | **0.22** | 0.23 | 0.25 | 0.26 | 0.29 | 0.28 |
| cond-mat | **0.29** | **0.29** | 0.33 | 0.51 | 0.37 | 0.46 |
| astro-ph | **0.31** | 0.33 | 0.31 | 0.48 | 0.47 | 0.50 |
| Lesmis | **0.20** | 0.41 | 0.37 | 0.32 | 0.38 | 0.48 |
| Celegansneural | **0.25** | 0.35 | 0.38 | 0.44 | 0.52 | 0.58 |
| Polbooks | **0.08** | 0.19 | 0.24 | 0.25 | 0.12 | 0.43 |

(جدول-۸) مقادیر ضریب خوشه‌بندی به‌دست‌آمده برای روش پیشنهادی و مقایسه آن با میانگین ضریب خوشه‌بندی شبکه اصلی و الگوریتم‌های دیگر با نرخ نمونه‌گیری بیست درصد

**(Table-8): The obtained clustering coefficient values for the proposed method and compared it with clustering coefficient of main networks and compared algorithms at a sampling rate of 20%**

| | ضریب خوشه‌بندی شبکه اصلی | روش پیشنهادی | الگوریتمCSN | الگوریتمNR | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Netscience** | 0.637 | 0.532 | 0.467 | **0.549** | 0.352 | 0.489 | 0.321 |
| **Internet_Routers-22july6** | 0.351 | **0.239** | 0.231 | 0.208 | 0.138 | 0.158 | 0.001 |





[ DOI: 10.29252/jsdp.17.3.55 ]

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| **Karate** | 0.571 | **0.351** | 0.334 | 0.299 | 0.229 | 0.229 | 0.22 |
| **Hep-th** | 0.441 | **0.398** | 0.301 | 0.275 | 0.175 | 0.255 | 0.072 |
| **Football** | 0.403 | **0.373** | 0.223 | 0.181 | 0.311 | 0.211 | 0.166 |
| **Dolphins** | 0.259 | 0.216 | **0.229** | 0.113 | 0.103 | 0.127 | 0.207 |
| **Cond-mat** | 0.620 | 0.489 | **0.553** | 0.375 | 0.285 | 0.435 | 0.353 |
| **Astro-ph** | 0.638 | **0.541** | 0.467 | 0.521 | 0.431 | 0.391 | 0.499 |
| **Lesmis** | 0.573 | **0.492** | 0.453 | 0.438 | 0.318 | 0.418 | 0.435 |
| **Celegansneural** | 0.292 | **0.251** | 0.212 | 0.195 | 0.222 | 0.196 | 0.189 |
| **Polbooks** | 0.487 | **0.369** | 0.307 | 0.327 | 0.277 | 0.247 | 0.149 |

**(جدول-۹): میانگین ضریب خوشه‌بندی به‌دست‌آمده برای روش پیشنهادی و مقایسه آن با شبکه اصلی و الگوریتم‌های دیگر با نرخ نمونه‌گیری سی درصد**

**(Table-9): The obtained clustering coefficient values for the proposed method and compared it with clustering coefficient of main networks and compared algorithms at a sampling rate of 30%**

| | ضریب خوشه‌بندی شبکه اصلی | روش پیشنهادی | الگوریتمCSN | الگوریتمNR | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Netscience** | 0.637 | **0.562** | 0.497 | 0.551 | 0.386 | 0.509 | 0.351 |
| **Internet_Routers-22july6** | 0.351 | **0.289** | 0.261 | 0.248 | 0.178 | 0.218 | 0.141 |
| **Karate** | 0.571 | **0.481** | 0.473 | 0.229 | 0.259 | 0.269 | 0.28 |
| **Hep-th** | 0.441 | **0.378** | 0.371 | 0.315 | 0.235 | 0.275 | 0.122 |
| **Football** | 0.403 | **0.383** | 0.253 | 0.231 | 0.351 | 0.261 | 0.216 |
| **Dolphins** | 0.259 | **0.213** | 0.189 | 0.127 | 0.162 | 0.167 | 0.227 |
| **Cond-mat** | 0.620 | 0.549 | **0.583** | 0.435 | 0.345 | 0.375 | 0.393 |
| **Astro-ph** | 0.638 | 0.571 | 0.497 | 0.551 | 0.471 | 0.411 | 0.529 |
| **Lesmis** | 0.573 | 0.492 | **0.523** | 0.478 | 0.368 | 0.468 | 0.465 |
| **Celegansneural** | 0.292 | 0.281 | 0.202 | 0.185 | **0.282** | 0.19 | 0.239 |
| **Polbooks** | 0.487 | 0.399 | 0.337 | 0.357 | 0.197 | 0.277 | 0.209 |

**(جدول-۱۰): میانگین ضریب خوشه‌بندی به‌دست‌آمده برای روش پیشنهادی و مقایسه آن با شبکه اصلی و الگوریتم‌های دیگر با نرخ نمونه‌گیری پنجاه درصد**

**(Table-10): The obtained clustering coefficient values for the proposed method and compared it with clustering coefficient of main networks and compared algorithms at a sampling rate of 50%**

| | ضریب خوشه‌بندی شبکه اصلی | روش پیشنهادی | الگوریتمCSN | الگوریتمNR | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Netscience** | 0.637 | **0.623** | 0.547 | 0.57 | 0.43 | 0.569 | 0.401 |
| **Internet_Routers-22july6** | 0.351 | **0.329** | 0.321 | 0.308 | 0.238 | 0.258 | 0.101 |
| **Karate** | 0.571 | **0.531** | 0.530 | 0.289 | 0.319 | 0.329 | 0.31 |
| **Hep-th** | 0.441 | **0.428** | 0.401 | 0.375 | 0.265 | 0.335 | 0.172 |
| **Football** | 0.403 | **0.413** | 0.303 | 0.261 | 0.391 | 0.311 | 0.256 |
| **Dolphins** | 0.259 | **0.263** | 0.229 | 0.267 | 0.247 | 0.217 | 0.287 |
| **Cond-mat** | 0.620 | 0.589 | **0.633** | 0.475 | 0.375 | 0.415 | 0.433 |
| **Astro-ph** | 0.638 | **0.621** | 0.557 | 0.611 | 0.521 | 0.471 | 0.579 |
| **Lesmis** | 0.573 | 0.522 | **0.545** | 0.527 | 0.406 | 0.508 | 0.525 |
| **Celegansneural** | 0.292 | **0.321** | 0.262 | 0.235 | 0.317 | 0.248 | 0.269 |
| **Polbooks** | 0.487 | **0.459** | 0.387 | 0.417 | 0.257 | 0.327 | 0.239 |



[ DOI: 10.29252/jsdp.17.3.55 ]

نتایج حاصل از جدول (۱۰) نیز نشان می‌دهد که کارآیی الگوریتم پیشنهادی در حفظ ضریب خوشه‌بندی در نمونه به‌دست‌آمده با نمونه‌گیری پنجاه درصد نسبت به دیگر روش‌های مقایسه‌شده در نُه شبکه از یازده شبکه آزمایش‌شده بالا هست. در بین روش‌های دیگر نیز الگوریتم CSN بهترین عملکرد را در دو شبکه Cond-mat و Lesmis دارد. تحلیل نتایج به‌دست‌آمده از هر سه جدول (۸، ۹ و ۱۰) ثابت می‌کند. که روش پیشنهادی از استحکام بسیار خوبی در یافتن نمونه‌ها برخوردار است؛ زیرا نتایج نشان می‌دهد که با افزایش نمونه‌گیری از بیست درصد به سی درصد و نیز از سی درصد به پنجاه درصد، روش پیشنهادی به‌قطعیت بهترین میزان حفظ و بهبود ضریب خوشه‌بندی را نسبت به بقیه روش‌ها در همه مجموعه‌داده‌ها به‌استثنای Lesmis دارد. در مجموعه‌داده Lesmis نیز، برخلاف این‌که در نمونه‌گیری پنجاه درصد، روش CSN عملکرد بهتری داشته ولی روش پیشنهادی حدود ۰/۰۳ بهبود نسبت به نمونه‌گیری سی درصد دارد؛ درحالی‌که در روش CSN، میزان بهبود ۰/۲۵ را نسبت به نمونه‌گیری سی درصد مشاهده می‌کنیم. به‌عبارتی، میزان نرخ خطا در روش پیشنهادی در پایین‌ترین سطح نسبت به روش‌های دیگر قرار دارد. می‌توان گفت که الگوریتم پیشنهادی منجر به تولید نمونه‌شبکه با کیفیت نزدیک به کیفیت شبکه اصلی شده، که به‌طورتقریبی تمام ویژگی‌های شبکه اصلی در شبکه نمونه‌گیری حفظ شده است.

**(جدول-۱۱): تعداد جوامع به‌دست‌آمده با الگوریتم LP-LPA در نمونه‌های حاصله با روش پیشنهادی و الگوریتم‌های دیگر (با نمونه‌گیری۳۰٪)**

**(Table-11): The number of discovered communities using LP-LPA for obtained sampling by the proposed method and the other compared algorithms at a sampling rate of 30%**

| | تعداد جوامع شبکه اصلی | روش پیشنهادی | CSNالگوریتم | NRالگوریتم | روش نمونه‌برداری TIES | نمونه‌برداری با الگوریتم ژنتیک | روش نمونه‌برداری تصادفی |
|---|---|---|---|---|---|---|---|
| **Karate** | 2 | **2** | **2** | 3 | 3 | 3±1 | 4±1 |
| **Dolphins** | 2 | **3** | 5 | 4 | 4 | 5±2 | 7±2 |
| **Football** | 12 | 9 | **10** | 7 | 6 | 13±7 | 15±5 |
| **Polbook** | 3 | **4** | 5 | 6 | **4** | 5+1 | 5+2 |
| **Lesmis** | 2 | **2** | 4 | 4 | 7 | 8±4 | 12±3 |

## ۴-۴- آنالیز براساس حفظ ساختار انجمن‌ها

در این قسمت با استفاده از فقط چهار شبکه واقعی Karate, Dolphins, Lesmis و Polbook تعداد جوامع موجود را در نمونه‌های به‌دست‌آمده در نمونه‌گیری باهم مقایسه کردیم زیرا در مابقی مجموعه‌دادگان، تعداد جوامع نامشخص هست. در همه روش‌های نمونه‌گیری نیز از الگوریتم LP-LPA [26] برای یافتن جوامع در شرایط به‌طورکامل یکسان استفاده کردیم. همان‌طوری‌که در جدول (۱۱) مشاهده می‌کنید، نمونه‌شبکهٔ به‌دست‌آمده از روش پیشنهادی علاوه‌بر این‌که ویژگی‌های توپولوژیکی و مدل شبکه اصلی یعنی مقیاسِ آزاد[1] بودن را حفظ کرده، بلکه ویژگی حفظ ساختار انجمنی و توزیع قانون توانی را نیز به‌طورکامل دارد؛ لذا کارایی الگوریتم پیشنهادی در حفظ مدل شبکه اصلی و انعکاس آن در نمونه به‌دست‌آمده نسبت به روش‌های دیگر به‌طورکامل مشهود هست. به‌استثنای شبکه Football، در مابقی مجموعه‌داده‌ها، روش پیشنهادی بهترین عملکرد را در حفظ ویژگی ساختار انجمنی دارد؛ ولی در Football، روش CSN عملکرد به‌نسبه بهتری دارد. روش‌هایی مثل نمونه‌گیری با الگوریتم ژنتیک و روش تصادفی، با مشکل ناپایداری نیز مواجه هستند؛ زیرا الگوریتم ژنتیک نیز انتخاب تصادفی داشته و در آنها برخی از گره‌ها در مؤلفه‌های به‌طورکامل مستقل و بدون ارتباط با یکدیگر هستند. برای مثال در مجموعه‌داده Lesmis که در شکل (۱) نشان داده شده است، می‌توانید نتایج روش ناپایداری مثل نمونه‌برداری مبتنی بر الگوریتم ژنتیک (شکل ۲) را با روش پیشنهادی (شکل ۳) مقایسه کنید. واضح است که روش پیشنهادی توانسته است به‌درستی هر دو جامعه را در نمونه حاصل، حفظ کند؛ ولی روشی مثل الگوریتم ژنتیک چنین ویژگی بارزی را ندارد و همان‌طور که مشاهده می‌کنید، دارای چندین جامعه تک‌گرهی است.

گفتنی است که نمونه‌برداری شبکه‌های واقعی به‌دلیل سعی در حفظ ویژگی‌های توپولوژیکی، به‌هیچ‌وجه قابل استفاده برای شبکه‌های تصادفی نیست. توجه داشته باشیم

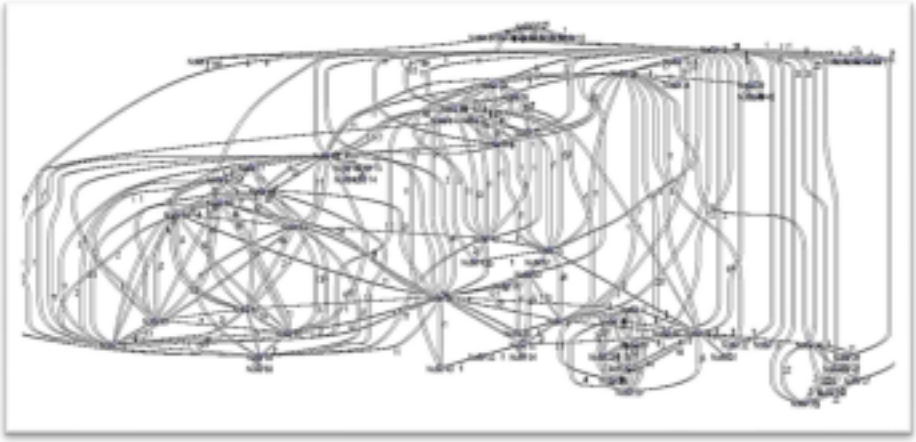
[1] Scale Free





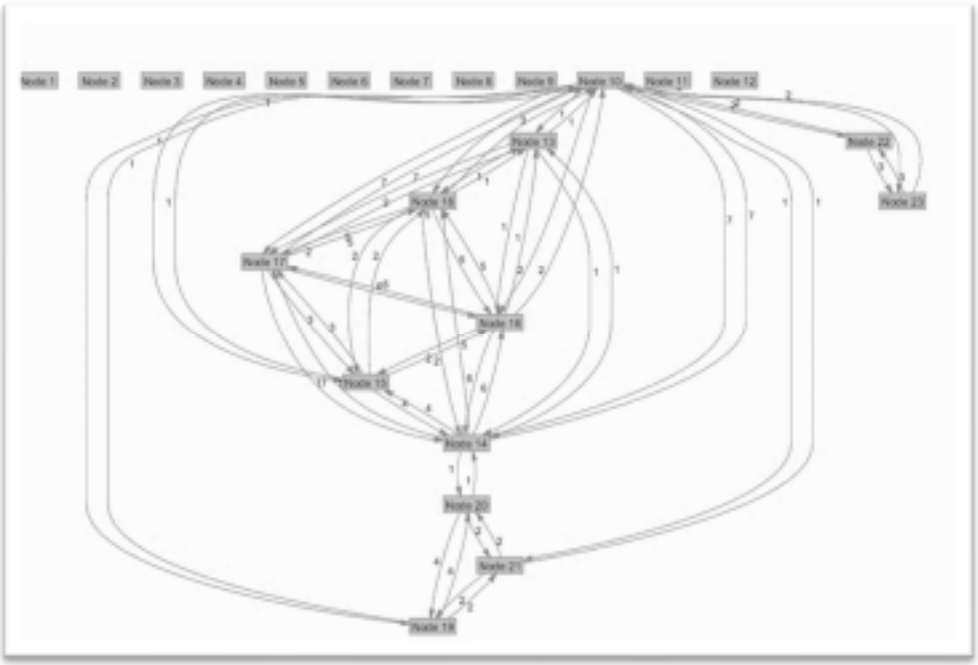
[ DOI: 10.29252/jsdp.17.3.55 ]

که تأکید ما در روش پیشنهادی، حفظ گره‌های مهم و استراتژیک در شبکه و درنتیجه پایداری شبکه حاصل است که این نوع گره‌ها به‌طورمعمول پایین‌ترین ضریب خوشه‌بندی را نسبت به گره‌های معمولی دارند و به‌طورمعمول یا تراکم جوامع را حفظ می‌کنند و یا باعث جلوگیری از متلاشی‌شدن شبکه و قطع ارتباط بین جوامع می‌شوند؛ در حالی‌که هیچ یک از موارد یادشده در گراف‌های تصادفی اهمیت ندارد؛ بنابراین باید گفت که این روش فقط برای شبکه‌های پیچیده مثل شبکه‌های اجتماعی مناسب است.



**(شکل-۱): شبکه‌اجتماعی Lesmis (شبکه‌ای با دو انجمن در حالت بدون نمونه‌گیری)**

**(Figure-1): Lesmis Social Network (a network with two communities in original network)**



**(شکل-۲): نمونه شبکه Lesmis به‌دست‌آمده با استفاده از روش نمونه‌گیری ژنتیک**

**(Figure-2): The obtained sample from Lesmis Social Network using genetic-based sampling algorithm**



[ DOI: 10.29252/jsdp.17.3.55 ]

**(شکل-۳): نمونه شبکه Lesmis به‌دست‌آمده با استفاده از روش نمونه‌گیری پیشنهادی**
**(Figure-3): The obtained sample from Lesmis Social Network using proposed sampling algorithm**

## ۵- نتیجه‌گیری

در این مقاله برای رفع مشکلات الگوریتم‌های نمونه‌گیری قبلی مانند، عدم توجه به حفظ گره‌های مهم، عدم حفظ ساختار گراف اصلی، عدم حفظ توزیع قانون توانی، و عدم حفظ ویژگی‌های توپولوژیکی شبکه اصلی در نمونه‌شبکهٔ به‌دست‌آمده در شبکه‌های اجتماعی و سایر شبکه‌های پیچیده، یک الگوریتم جدیدی با استفاده از یک رابطه وزن‌دهی محلی و به‌کارگیری آن در الگوریتم کلونی زنبورعسل مبتنی بر انتخاب آگاهانه نمونه‌ها برای بهبود دقت نمونه‌گیری ارائه داده شد. در روش پیشنهادی پس از رتبه‌بندی و تشخیص گره‌های مهم شبکه اجتماعی همانند گره هسته و هاب و یا پل‌ها، بر اساس رابطه وزن‌دهی محلی تلاش شد تا گره‌های بااهمیت نسبت به گره‌های پیرامونی[1] شبکه تا حد ممکن در نمونه‌شبکهٔ تولیدشده حضور داشته باشند تا نمونه حاصل از لحاظ کیفیت به گراف شبکه اصلی مورد نظر تبه‌طورتقریبی مشابهت داشته باشد؛ زیرا حفظ ساختار جامعه‌ها و ویژگی‌های اتصالی مؤلفهٔ هم‌بند عظیم در گرو انتخاب بخش زیادی از همین گره‌های با اهمیت هست.

برای اثبات کارایی روش پیشنهادی، آن را بر روی مجموعه‌دادگان استاندارد شبکه اجتماعی و پیچیده، از جنبه‌های مختلف مانند میانگین ضریب خوشه‌بندی، تحلیل توزیع درجه بر حسب مقدار KS، پارامترهای $\alpha$ و $X_{min}$، و حفظ ساختار انجمن‌ها مورد آزمایش قرار دادیم. در تمام ارزیابی‌های صورت‌گرفته، روش پیشنهادی بهترین عملکرد را نسبت به سایر روش‌ها دارد. به‌عنوان مثال، از نظر توزیع درجه بر حسب مقدار KS، و پارامترهای $\alpha$ و $X_{min}$، روش پیشنهادی کارایی بالای نمونه‌گیری را نسبت به روش‌های بر روی مجموعه‌داده‌های مختلفی از لحاظ معیار خطای نسبی نشان داد. در سایر معیارها مانند حفظ ضریب خوشه‌بندی و ساختارهای انجمنی در نمونه حاصل نیز نتایج به‌دست‌آمده در جداول (۸ تا ۱۱) نشان داد که روش پیشنهادی به‌دلیل اهمیت‌دادن به گره‌های کلیدی مثل گره‌های هسته و هاب تا حد زیادی مانع از به‌هم‌خوردن توپولوژی شبکه می‌شود و به‌طبع آن تعداد انجمن‌ها و ضریب خوشه‌بندی شبکه اصلی و شبکه نمونه‌برداری‌شده مشابهت بالایی را در روش پیشنهادی نشان می‌دهد.

یک محدودیت مهم روش پیشنهادی، عدم توجه به گره‌های پل با درجهٔ پایین و همسایه‌های درجهٔ پایین هست، در حالی‌که این نوع پل‌ها ممکن است دو بخش مهم پرتراکم گراف را به‌هم متصل می کنند؛ بنابراین، رفع این محدودیت می‌تواند به‌عنوان یک کار پژوهشی جدید باشد؛ همچنین به‌عنوان پیشنهادی برای کارهای آینده، یک کار جدید، استفاده از الگوریتم تشخیص جوامع هم‌پوشان در فرایند نمونه‌برداری هست. چنانچه این گام در ابتدا اجرا شود می‌توان با محاسبه ارزش هر جامعه و نیز درنظرگرفتن موقعیت استراتژیکی گره‌های هم‌پوشان و نیز اهمیت گره‌های پخش‌کننده[2] یک نمونه با کیفیت و کارا برای شبکه اصلی ایجاد کرد. در گام دوم می‌توان نمونه‌برداری را از جوامع مهم و از گره‌های مهم درون جامعه و گره‌های هم‌پوشان شروع کرد و تا جوامع ضعیف ادامه داد.

[1] periphery

[2] Spreader





[ DOI: 10.29252/jsdp.17.3.55 ]

## 6- References ۶- مراجع


[1] E. Katz, P. F. Lazarsfeld, and E. Roper, *Personal influence: The part played by people in the flow of mass communications*. Routledge, 2017.

[2] N. B. Ellison, J. Vitak, R. Gray, and C. Lampe, "Cultivating social resources on social network sites: Facebook relationship maintenance behaviors and their role in social capital processes," *Journal of Computer-Mediated Communication,* vol. 19, no. 4, pp. 855-870, 2014.

[3] M. Irani and M. Haghighi, "The Impact of Social Networks on the Internet Business Sustainability (With Emphasis on the Intermediary Role of Entrepreneurial Purpose of Online Branches of Mellat Bank's Portal)," *Journal of Information Technology Management,* vol. 5, no. 4, pp. 23-46, 2013.

[4] M. Emirbayer and J. Goodwin, "Network analysis, culture, and the problem of agency," *American journal of sociology,* vol. 99, no. 6, pp. 1411-1454, 1994.

[5] J. Tang, J. Sun, C. Wang, and Z. Yang, "Social influence analysis in large-scale networks," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2009, pp. 807-816: ACM .

[6] J. Török, Y. Murase, H.-H. Jo, J. Kertész, and K. Kaski, "What big data tells: sampling the social network by communication channels," *Physical Review E,* vol. 94, no. 5, pp. 052319, 2016.

[7] M. Papagelis, G. Das, and N. Koudas, "Sampling online social networks," *IEEE Transactions on knowledge and data engineering,* vol. 25, no. 3, pp. 662-676, 2013.

[8] K. Dempsey, K. Duraisamy, H. Ali, and S. Bhowmick, "A parallel graph sampling algorithm for analyzing gene correlation networks," *Procedia Computer Science,* vol. 4, pp. 136-145, 2011.

[9] J. Leskovec and C. Faloutsos, "Sampling from large graphs," in *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2006, pp. 631-636: ACM.

[10] S. Yoon, S. Lee, S.-H. Yook, and Y. Kim, "Statistical properties of sampled networks by random walks," *Physical Review E,* vol. 75, no. 4, pp. 046114, 2007.

[11] S.-H. Yoon, K.-N. Kim, J. Hong, S.-W. Kim, and S. Park, "A community-based sampling method using DPL for online social networks," *Information Sciences,* vol. 306, pp. 53-69, 2015.

[12] E. M. Airoldi, X. Bai, and K. M. Carley, "Network sampling and classification: An investigation of network model representations," *Decision support systems,* vol. 51, no. 3, pp. 506-518, 2011.

[13] P. Krömer and J. Platoš, "Genetic algorithm for sampling from scale-free data and networks," in *Proceedings of the 2014 annual conference on genetic and evolutionary computation*, 2014, pp. 793-800: ACM.

[14] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines," *The journal of chemical physics,* vol. 21, no. 6, pp. 1087-1092, 1953.

[15] L. A. Goodman, "Snowball sampling," *The annals of mathematical statistics,* pp. 148-170, 1961.

[16] L. Lovász, "Random walks on graphs: A survey," *Combinatorics, Paul erdos is eighty,* vol. 2, no. 1, pp. 1-46, 1993.

[17] D. D. Heckathorn, "Respondent-driven sampling: a new approach to the study of hidden populations," *Social problems,* vol. 44, no. 2, pp. 174-199, 1997.

[18] S. Ye, J. Lang, and F. Wu, "Crawling online social graphs," in *2010 12th International Asia-Pacific Web Conference:IEEEP*, pp. 236-242, 2010.

[19] A. Rezvanian and M. R. Meybodi, "Sampling social networks using shortest paths," *Physica A: Statistical Mechanics and its Applications,* vol. 424, pp. 254-268, 2015.

[20] A. Sevilla, A. Mozo, and A. F. Anta, "Node sampling using random centrifugal walks," *Journal of Computational Science,* vol. 11, pp. 34-45, 2015.

[21] C. Tong, Y. Lian, J. Niu, Z. Xie, and Y. Zhang, "A novel green algorithm for sampling complex networks," *Journal of Network and Computer Applications,* vol. 59, pp. 55-62, 2016.

[22] N. Ahmed, J. Neville, and R. R. Kompella, "Network sampling via edge-based node selection with graph induction," 2011.

[23] K. Brádler, P.-L. Dallaire-Demers, P. Rebentrost, D. Su, and C. Weedbrook, "Gaussian boson sampling for perfect matchings of arbitrary graphs," *Physical Review A,* vol. 98, no. 3, pp. 032310, 09/10/ 2018.

[24] J. Zhao, P. Wang, J. C. S. Lui, D. Towsley, and X. Guan, "Sampling online social networks by random walk with indirect jumps," *Data Mining and Knowledge Discovery,* journal article vol. 33, no. 1, pp. 24-57, January 01. 2019.

[25] Y. Xie, S. Chang, Z. Zhang, M. Zhang, and L. Yang, "Efficient sampling of complex network with modified random walk strategies," *Physica*




[ DOI: 10.29252/jsdp.17.3.55 ]

*A: Statistical Mechanics and its Applications,* vol. 492, pp. 57-64, 2018.

[26] K. Berahmand and A. Bouyer, "LP-LPA: A link influence-based label propagation algorithm for discovering community structures in networks," *International Journal of Modern Physics B,* vol. 32, no. 06, pp. 1850062, 2018.

**عسگرعلی بویر** دانشیار گروه مهندسی رایانه دانشکده فناوری اطلاعات و مهندسی رایانه در دانشگاه شهید مدنی آذربایجان بوده که در سال ۱۳۹۰، مقطع دکترا را در رشته مهندسی رایانه از دانشگاه صنعتی مالزی(UTM) اخذ کرده است. وی در حال حاضر مدیر مسئول آزمایشگاه پژوهشی رایانش توزیعی و کلان‌داده است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: تحلیل شبکه‌های پیچیده/ اجتماعی، رایانش توزیعی مبتنی بر رایانش ابری/ گرید و کاربرد آن در مسائل کلان‌داده، داده‌کاوی و کاربردهای آن در صنعت و پزشکی، و امنیت شبکه‌های رایانه‌ای و محیط‌های توزیعی.

نشانی رایانامه ایشان عبارت است از:

**a.bouyer@azaruniv.ac.ir**

**سمیه نوروزی** تحصیلات کارشناسی ارشد خود را در رشته مهندسی رایانه از دانشگاه آزاد اسلامی واحد میاندوآب در سال ۱۳۹۵ اخذ کرده است. زمینه‌های پژوهشی مورد علاقه ایشان عبارتند از: تحلیل شبکه‌های پیچیده/ اجتماعی و داده‌کاوی کاربردی است.

نشانی رایانامه ایشان عبارت است از:

**someiyenorozi2014@gmail.com**



[ DOI: 10.29252/jsdp.17.3.55 ]