

بخش اول

سؤال ۱. Gini Index چیست؟ تعریف آن را با Lorenz Curve توضیح دهید. کجاها از این معیار استفاده می‌شود؟

پاسخ صحیح شامل نکات زیر است:

۱. تعریف Gini Index

Gini Index معیاری برای سنجش برابری توزیع یک منبع میان یک مجموعه از موجودیت‌ها است؛ مقداری بین 0 (توزیع کاملاً برابر) تا ۱ (تمرکز کامل منبع در دست یک یا چند موجودیت).

۲. توضیح Lorenz Curve

○ Lorenz Curve منحنی‌ای است که درصد تجمعی جمعیت (محور افقی) را در برابر درصد تجمعی منبع (محور عمودی) رسم می‌کند.

○ Gini Index برابر دو برابر ناحیه بین منحنی لورنز و خط برابری کامل (قطر مربع) است.

۳. کاربردها

○ اقتصاد: سنجش نابرابری ثروت یا درآمد.

○ پردازش سیگنال: سنجش فشردگی سیگنال‌ها.

○ گراف‌های شبکه‌ای: تحلیل توزیع درجات نودها یا وزن یال‌ها در گراف‌ها (موضوع مقاله).

سؤال ۲. چرایی و چگونگی استفاده از Gini Index در تعیین sparsity یک گراف را شرح دهید.

پاسخ صحیح شامل نکات زیر است:

۱. دلایل استفاده

○ Edge density تنها یک نسبت ساده است و تنوع بین درجات نودها را منعکس نمی‌کند.

○ Gini Index قادر است تفاوت در توزیع درجات را نشان دهد، و بنابراین یک دید دقیق‌تر نسبت به

sparsity می‌دهد.

۲. روش استفاده

○ با استفاده از بردار درجات نودها $a = [a_1, \dots, a_n]$ ابتدا درجات مرتب می‌شوند.

○ سپس با کمک Lorenz Curve و محاسبه مساحت زیر آن، Gini Index به دست می‌آید.

○ برای اندازه‌گیری sparsity، نویسندگان شاخص جدیدی به نام Sparsity Index (SI) معرفی می‌کنند که

نسخه تعمیم‌یافته Gini است با مخرجی برابر حداکثر درجات ممکن $n(n-1)$:

۳. تفاوت با edge density و مزیت‌های SI

○ SI می‌تواند برای گراف‌های با edge density یکسان ولی توزیع درجات متفاوت، تمایز قائل شود.

- SI اطلاعات غنی‌تری ارائه می‌دهد و برای گراف‌های وزندار و نامتقارن نیز قابل تعمیم است.

سؤال ۳) Sparsity. یک Complete graph و Regular cycle را با Gini index به دست آورده و مزایا و معایب آن را تحلیل نمایید.

پاسخ صحیح شامل نکات زیر است:

۱. Complete Graph

- در Complete Graph همه نودها دارای درجه $n-1$ هستند.
- توزیع درجات کاملاً برابر است $\Rightarrow \text{Gini Index} = 0 \Rightarrow \text{Sparsity} = 0$
- مطابق انتظار، چون هیچ گونه sparsity وجود ندارد.

۲. Regular Cycle

- هر نود درجه ۲ دارد \Rightarrow توزیع برابر $\Rightarrow \text{Gini Index} = 0$
- اما چون حداکثر درجات ممکن بسیار بیشتر است، گراف عملاً sparse است.
- در این حالت Gini به تنهایی این sparsity را منعکس نمی‌کند.

۳. تحلیل مزایا و معایب

- مزایا Gini Index: سادگی محاسبه دارد و نابرابری را به خوبی نمایش می‌دهد.
- معایب: در گراف‌هایی با درجات برابر مانند Regular Cycle ممکن است sparsity واقعی را منعکس نکند.
- این ضعف با تعریف SI رفع شده که درجات را نسبت به یک مقدار ثابت (مثلاً حداکثر مجموع درجات ممکن) نرمال می‌کند.

سؤال ۴. پیچیدگی محاسباتی Sparsity Index را تحلیل نمایید.

پاسخ صحیح شامل نکات زیر است:

۱. مراحل محاسبه SI

- مرتب‌سازی بردار درجات نودها.
- محاسبه مساحت زیر Lorenz Curve با تقریب قطعه‌ای.

- استفاده از فرمول SI که بر پایه Gini است ولی مخرج آن بر اساس درجات بالقوه $(n-1)$ تنظیم شده.

۲. تحلیل پیچیدگی

- مرتب‌سازی نیاز به $O(n \log n)$ دارد (اصلی‌ترین عامل)
- باقی مراحل شامل جمع و اعمال فرمول‌های خطی هستند $O(n)$
- در مجموع پیچیدگی زمانی $O(n \log n)$

بخش دوم

سؤال ۱) روش‌های مبتنی بر LLM در مقایسه با مدل‌های کلاسیک در چه جنبه‌هایی برتری دارند و در چه جنبه‌هایی محدودیت دارند؟

برتری‌ها:

- درک زمینه‌ای پیچیده: مدل‌های زبانی بزرگ (LLM) می‌توانند اطلاعات زمینه‌ای مانند جنسیت، سن، نژاد، مذهب، وابستگی سیاسی و علایق را پردازش و در تولید روابط اجتماعی لحاظ کنند.
- توانایی ایجاد ساختارهای شبکه‌ای پیچیده: برخلاف مدل‌های کلاسیک (مثل SBM یا ERG)، LLM ها می‌توانند ویژگی‌هایی مانند خوشه‌بندی، توزیع درجه سنگین‌دم (heavy-tailed) و ویژگی‌های مرتبط با ساختار جوامع انسانی را بهتر تقلید کنند.
- تطبیق پذیری بالا: LLM ها به راحتی برای دامنه‌ها و ویژگی‌های جدید بدون نیاز به تعریف رسمی مدل یا بازآموزی قابل استفاده هستند.

محدودیت‌ها:

- سوگیری و بیش‌برآوردی (Overestimation): یکی از یافته‌های اصلی مقاله این است که LLM ها تمایل دارند میزان هم‌گروهی سیاسی (Political Homophily) را بیشتر از حد واقعی تخمین بزنند.

- شفافیت پایین و فقدان تبیین پذیری: برخلاف مدل‌های آماری، فرآیندهای درونی LLM ها تبیین پذیر نیستند و نمی‌توان دقیقاً گفت چرا یک رابطه خاص ایجاد شده.
- پایداری پایین در روش‌های غیر ترکیبی: شبکه‌هایی که به صورت global یا local تولید می‌شوند، اغلب ساختارهایی ناسازگار با شبکه‌های واقعی دارند.

سؤال ۲) دو روش اصلی "Global" و "Local" برای تولید شبکه‌های اجتماعی در مقاله توضیح داده شده‌اند. کدام روش واقعی‌تر است و چرا؟

روش Local نسبت به روش Global ساختارهای شبکه‌ای واقعی‌تری تولید می‌کند. در روش Global، کل ساختار شبکه یکباره توسط مدل تولید می‌شود و این روش منجر به شبکه‌هایی با ویژگی‌های غیرواقعی (مثل چگالی بیش از حد بالا یا پایین، یا عدم وجود خوشه‌بندی طبیعی) می‌شود. در روش Local، ارتباطات برای هر فرد جداگانه و مبتنی بر ویژگی‌های شخصیتی و زمینه‌ای ایجاد می‌شود. این روش منجر به شبکه‌هایی با درجه واقع‌گرایی بیشتر در ویژگی‌هایی مثل:

- توزیع درجه،
- چگالی شبکه،
- خوشه‌بندی،
- و ساختار محلی روابط می‌شود.

با این حال، حتی روش Local هم در بازتولید دقیق توزیع درجه شکست می‌خورد، که در سؤال ۳ به آن پرداخته می‌شود.

سؤال ۳) چگونه روش "Sequential" می‌تواند توزیع درجه (Degree Distribution) شبکه‌های واقعی را بهتر از روش "Local" مدل کند؟

در روش Sequential، مدل‌سازی روابط به صورت مرحله‌ای انجام می‌شود؛ یعنی هر گره جدید با توجه به گره‌های قبلی و روابطشان وارد شبکه می‌شود. این روش برخلاف Local، از تاریخچه ارتباطات قبلی آگاه است و می‌تواند در تولید ارتباطات جدید از الگوهای قبلی تبعیت کند (نوعی حافظه ضمنی).

به‌طور خاص، در روش Sequential

- احتمال اتصال به گره‌هایی که قبلاً Degree بالاتری داشته‌اند بیشتر است (الگوی مشابه Preferential Attachment).

- توزیع درجه ایجادشده شبیه به توزیع سنگین‌دم در شبکه‌های اجتماعی واقعی است.
- میزان خوشه‌بندی و ویژگی‌های محلی نیز حفظ می‌شوند.

بنابراین، Sequential نه تنها توزیع درجه واقع‌گراتری تولید می‌کند، بلکه ساختار اجتماعی واقعی‌تری نیز بازتولید می‌کند.

سؤال ۴) اگر بخواهید آزمایشی برای بررسی میزان واقعی بودن شبکه‌های اجتماعی تولید شده توسط LLM طراحی کنید، چه معیارهایی را اندازه می‌گیرید؟

برای ارزیابی میزان شباهت شبکه‌های تولیدشده به شبکه‌های واقعی، معیارهای زیر باید بررسی شوند:

۱. توزیع درجه: (Degree Distribution) آیا توزیع درجه سنگین‌دم است؟ آیا رفتار long-tail دارد؟
۲. چگالی شبکه: (Density) نسبت تعداد یال‌ها به حداکثر یال‌های ممکن.
۳. ضریب خوشه‌بندی: (Clustering Coefficient) میانگین احتمال تشکیل مثلث‌ها در گراف.
۴. طول متوسط مسیر: (Average Path Length) میانگین فاصله بین همه جفت گره‌ها.
۵. هم‌گروهی: (Homophily) مخصوصاً هم‌گروهی سیاسی؛ چقدر احتمال دارد افراد با ویژگی‌های مشابه با هم ارتباط داشته باشند؟
۶. Modularity یا ساختار جوامع: آیا شبکه به خوشه‌ها یا جوامع قابل شناسایی تقسیم می‌شود؟
۷. ویژگی‌های نودها: (Node-level Attributes) بررسی رابطه بین ویژگی‌هایی مثل سن، نژاد یا علاقه‌مندی‌ها با الگوهای اتصال.

همچنین می‌توان از مقایسه آماری با شبکه‌های واقعی (مثلاً با استفاده از Earth Mover Distance یا KL Divergence برای توزیع‌ها) نیز استفاده کرد.