



دانشگاه صنعتی امیرکبیر  
(پلی‌تکنیک تهران)  
دانشکده مهندسی کامپیوتر

پروژه کارشناسی  
گرایش کامپیوتر

توسعه سامانه ای برای رتبه بندی گره ها در شبکه های  
پیچیده از طریق شبکه های عصبی پیچشی

پایان نامه

نگارش  
نگین خیرمند

استاد راهنما  
دکتر مصطفی حقیر چهرقانی

۱۴۰۳ دی

بِسْمِ اللّٰهِ الرَّحْمٰنِ الرَّحِيْمِ

## صفحه فرم ارزیابی و تصویب پایان نامه- فرم تأیید اعضاء کمیته دفاع

در این صفحه فرم دفاع یا تایید و تصویب پایان نامه موسوم به فرم کمیته دفاع- موجود در پرونده آموزشی- را قرار دهید.

نکات مهم:

- نگارش پایان نامه/رساله باید به زبان فارسی و بر اساس آخرین نسخه دستورالعمل و راهنمای تدوین پایان نامه های دانشگاه صنعتی امیرکبیر باشد.(دستورالعمل و راهنمای حاضر)
- رنگ جلد پایان نامه/رساله چاپی کارشناسی، کارشناسی ارشد و دکترا باید به ترتیب مشکی، طوسی و سفید رنگ باشد.
- چاپ و صحافی پایان نامه/رساله بصورت پشت و رو(دورو) بلامانع است و انجام آن توصیه می شود.



دانشگاه صنعتی امیرکبیر  
(پلی‌تکنیک تهران)

به نام خدا

## تعهدنامه اصالت اثر

تاریخ: دی ۱۴۰۳

اینجانب نگین خیرمند متعهد می‌شوم که مطالب مندرج در این پایان‌نامه حاصل کار پژوهشی اینجانب تحت نظرارت و راهنمایی استادی دانشگاه صنعتی امیرکبیر بوده و به دستاوردهای دیگران که در این پژوهش از آنها استفاده شده است مطابق مقررات و روال متعارف ارجاع و در فهرست منابع و مأخذ ذکر گردیده است. این پایان‌نامه قبلاً برای احراز هیچ مدرک هم‌سطح یا بالاتر ارائه نگردیده است.

در صورت اثبات تخلف در هر زمان، مدرک تحصیلی صادر شده توسط دانشگاه از درجه اعتبار ساقط بوده و دانشگاه حق پیگیری قانونی خواهد داشت.

کلیه نتایج و حقوق حاصل از این پایان‌نامه متعلق به دانشگاه صنعتی امیرکبیر می‌باشد. هرگونه استفاده از نتایج علمی و عملی، واگذاری اطلاعات به دیگران یا چاپ و تکثیر، ترجمه و اقتباس از این پایان‌نامه بدون موافقت کتبی دانشگاه صنعتی امیرکبیر ممنوع است. نقل مطالب با ذکر مأخذ بلامانع است.

نگین خیرمند

امضا

## سپاسگزاری

از پدر، مادر و برادر عزیزم برای حمایت‌های بیدریغشان و از استاد راهنمای محترم، جناب آقای دکتر مصطفی حقیر چهرقانی، بابت راهنمایی‌ها و نظرات ارزشمندانه صمیمانه سپاسگزارم. همچنین می‌خواهم از جناب آقای دکتر علیرضا باقری که زحمت داوری این پروژه را بر عهده داشتند نیز نهایت تشکر را به جا آورم.

نگین خیرمند

۱۴۰۳

## چکیده

در این پژوهش، هدف اصلی پیاده‌سازی روشی نوین برای رتبه‌بندی گره‌های تأثیرگذار در شبکه‌های پیچیده با استفاده از شبکه‌های عصبی پیچشی (CNN) است. انتخاب گره‌های تأثیرگذار برای تحلیل و مدیریت شبکه‌های پیچیده اهمیت زیادی دارد، زیرا می‌تواند در بهبود کارایی الگوریتم‌ها و تصمیم‌گیری‌های شبکه مؤثر باشد. روش پیشنهادی در این تحقیق از شبکه‌های عصبی پیچشی برای استخراج ویژگی‌های گره‌ها و مدل‌سازی روابط پیچیده بین آن‌ها در شبکه‌های گرافی استفاده می‌کند. در ابتدا، شبکه پیچیده به صورت گراف مدل‌سازی شده و ویژگی‌های مختلف گره‌ها استخراج می‌شود. سپس، با استفاده از CNN، الگوهای پنهان و ساختار شبکه شناسایی شده و گره‌ها به ترتیب تأثیرگذاری رتبه‌بندی می‌شوند. نتایج آزمایش‌ها نشان می‌دهد که این روش قادر به شناسایی گره‌های کلیدی در شبکه‌ها با دقت و کارایی بالا است. این تحقیق راهکاری نوین برای تحلیل شبکه‌های پیچیده ارائه می‌دهد و ابزارهای پیشرفته‌ای برای تحلیل و تصمیم‌گیری در حوزه‌هایی مانند شبکه‌های اجتماعی، زیستی و اقتصادی فراهم می‌کند.

واژه‌های کلیدی:

شبکه‌های پیچیده، رتبه‌بندی گره‌های تأثیرگذار، شبکه‌های عصبی پیچشی، مدل‌سازی و استخراج ویژگی‌های شبکه‌های گرافی

# فهرست مطالب

عنوان	صفحة
۲ پیشینه موضوع	۷
۱-۲ مروری بر تحقیقات پیشین و مطالعات مرتبط	۸
۳ داده‌ها	۱۲
۱-۳ ساختار فصل	۱۳
۲-۳ تعاریف	۱۳
۱-۲-۳ مدل گراف باراباسی-آلبرت	۱۳
۱-۱-۲-۳ ویژگی‌های اصلی	۱۴
۲-۱-۲-۳ مزایا و محدودیت‌ها	۱۵
۲-۲-۳ مدل شبیه‌سازی SIR	۱۶
۱-۲-۲-۳ اجزای مدل	۱۶
۲-۲-۲-۳ آستانه انتشار و نظریه میدان میانگین	۱۷
۳-۲-۲-۳ شبیه‌سازی	۱۸
۴-۲-۲-۳ هدف از شبیه‌سازی	۱۹
۵-۲-۲-۳ فرمول تأثیرگذاری گره‌ها در مدل SIR	۱۹
۶-۲-۲-۳ مزایا و محدودیت‌ها	۱۹
۷-۲-۲-۳ فرآیند شبیه‌سازی	۲۰
۳-۲-۳ داکر	۲۰
۱-۳-۲-۳ ویژگی‌های کلیدی	۲۱
۳-۳ مجموعه‌ی داده‌ها	۲۲
۱-۳-۳ کار با داده	۲۲
۱-۱-۳-۳ انتخاب، ساخت یا جمع آوری مجموعه‌ی داده‌ها	۲۲
۲-۱-۳-۳ ساخت گراف‌های BA	۲۳
۳-۱-۳-۳ اصلاح و یکپارچه‌سازی داده‌های گرافی برای تحلیل	۲۶
۴-۱-۳-۳ تولید برچسب قدرت انتشار	۲۶
۵-۱-۳-۳ اهمیت برچسب قدرت انتشار و توزیع مقادیر آن‌ها	۲۶

۲۷	۶-۱-۳-۳ تاثیر پارامتر $\beta$ روی برچسب های تولیدی
۳۰	۷-۱-۳-۳ تاثیر دقت های مختلف در شبیه‌سازی
۳۱	۸-۱-۳-۳ تاثیر تعداد تکرار شبیه سازی
۳۲	۹-۱-۳-۳ تحلیل زمان اجرای ساخت برچسبها
۳۴	۱۰-۱-۳-۳ تسریع فرآیند شبیه‌سازی
۳۶	۱۱-۱-۳-۳ یکپارچه‌سازی و سازماندهی فایل‌های تولید شده توسط شبیه‌سازی
۳۸	۴ سامانه توسعه‌یافته
۳۹	۱-۴ شبکه عصبی پیچشی گرافی
۴۰	۱-۱-۴ معرفی مدل
۴۰	۲-۱-۴ مراحل اصلی الگوریتم
۴۰	۱-۲-۱-۴ استخراج ویژگی‌های گراف
۴۰	۲-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی مستقیم ( $W^{D1}$ )
۴۱	۳-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی تجمعی ( $W^{D2}$ )
۴۱	۴-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی تعمیم‌یافته ( $W^{D3}$ )
۴۱	۵-۲-۱-۴ ماتریس مرکزیت شاخص $H$ مستقیم همسایگی ( $W^{H1}$ )
۴۲	۶-۲-۱-۴ ماتریس مرکزیت شاخص $H$ تجمعی همسایگی ( $W^{H2}$ )
۴۲	۷-۲-۱-۴ ماتریس مرکزیت شاخص $H$ تعمیم‌یافته همسایگی ( $W^{H3}$ )
۴۲	۸-۲-۱-۴ ساخت ماتریس مجاورت
۴۳	۹-۲-۱-۴ ترکیب کانال‌ها و ساخت مجموعه کانال‌های ساختاری بازنمایی گره‌ها
۴۴	۱۰-۲-۱-۴ پیش‌بینی نهایی
۴۵	۳-۱-۴ پارامترهای مدل
۴۵	۴-۲-۴ سامانه توسعه‌یافته
۴۵	۱-۲-۴ توسعه رابط کاربری مبتنی بر وب
۴۶	۲-۲-۴ مزایای استفاده از کتابخانه Streamlit
۴۶	۳-۲-۴ صفحات رابط کاربری
۴۷	۴-۲-۴ پیش‌پردازش داده‌ها برای رابط کاربری

۴۹	.....	۵	ارزیابی
۵۰	.....	۱-۵	۱-۵ مقدمه
۵۰	.....	۲-۵	۲-۵ تعاریف معیارهای ارزیابی
۵۰	..... ضریب همبستگی Kendall's τ Correlation Coefficient کندال	۱-۲-۵	(Kendall's τ Correlation Coefficient) کندال
۵۱	..... Mean Average Precision (MAP)	۲-۲-۵	۲-۲-۵ ضرورت معیارهای ارزیابی انتخاب شده در ارزیابی
۵۳	..... ۱-۳-۲-۵ توانایی ارزیابی دقیق رتبه بندی ها	۳-۲-۵	۳-۲-۵ ضرورت معیارهای ارزیابی انتخاب شده در ارزیابی
۵۴	..... پوشش جنبه های مختلف دقت در شبکه های پیچیده	۲-۳-۲-۵	۲-۳-۲-۵ پوشش جنبه های مختلف دقت در شبکه های پیچیده
۵۴	..... ۳-۳-۲-۵ هماهنگی با اهداف مدل	۳-۳-۲-۵	۳-۳-۲-۵ هماهنگی با اهداف مدل
۵۴	..... ۴-۳-۲-۵ سازگاری با تحلیل عملکرد در مقیاس های مختلف	۴-۳-۲-۵	۴-۳-۲-۵ سازگاری با تحلیل عملکرد در مقیاس های مختلف
۵۵	..... ۵-۳-۲-۵ کاهش پیچیدگی محاسباتی	۵-۳-۲-۵	۵-۳-۲-۵ کاهش پیچیدگی محاسباتی
۵۵	..... ۶-۳-۲-۵ مقایسه و تحلیل جامع عملکرد مدل	۶-۳-۲-۵	۶-۳-۲-۵ مقایسه و تحلیل جامع عملکرد مدل
۵۵	..... ۳-۵ تنظیمات اولیه و پیکربندی آزمایش ها و پارامترها	۳-۵	۳-۵ تنظیمات اولیه و پیکربندی آزمایش ها و پارامترها
۵۵	..... ۱-۳-۵ تنظیمات بهینه سازی و پارامترهای آموزش	۱-۳-۵	۱-۳-۵ تنظیمات بهینه سازی و پارامترهای آموزش
۵۶	..... ۲-۳-۵ مشخصات فنی اجرای آزمایش ها	۲-۳-۵	۲-۳-۵ مشخصات فنی اجرای آزمایش ها
۵۶	..... ۴-۵ تحلیل عملکرد	۴-۵	۴-۵ تحلیل عملکرد
۵۶	..... ۱-۴-۵ دقت مدل در شناسایی گره های برتر	۱-۴-۵	۱-۴-۵ دقت مدل در شناسایی گره های برتر
۵۷	..... ۱-۱-۴-۵ بررسی توزیع مقادیر SIR	۱-۱-۴-۵	۱-۱-۴-۵ بررسی توزیع مقادیر SIR
۵۸	..... ۲-۴-۵ مقایسه عملکرد مدل های پیشنهادی	۲-۴-۵	۲-۴-۵ مقایسه عملکرد مدل های پیشنهادی
۵۹	..... ۱-۲-۴-۵ تحلیل دقت مدل برای یک تنظیم خاص	۱-۲-۴-۵	۱-۲-۴-۵ تحلیل دقت مدل برای یک تنظیم خاص
۶۰	..... ۲-۲-۴-۵ جمع بندی	۲-۲-۴-۵	۲-۲-۴-۵ جمع بندی
۶۰	..... ۳-۴-۵ معرفی الگوریتم های پایه	۳-۴-۵	۳-۴-۵ معرفی الگوریتم های پایه
۶۱	..... ۱-۳-۴-۵ مقایسه با الگوریتم های پایه	۱-۳-۴-۵	۱-۳-۴-۵ مقایسه با الگوریتم های پایه
۶۲	..... ۴-۴-۵ مزایا و نقاط قوت مدل پیشنهادی	۴-۴-۵	۴-۴-۵ مزایا و نقاط قوت مدل پیشنهادی
۶۲	..... ۱-۴-۴-۵ کارایی بالا در شناسایی گره های تأثیرگذار	۱-۴-۴-۵	۱-۴-۴-۵ کارایی بالا در شناسایی گره های تأثیرگذار
۶۲	..... ۲-۴-۴-۵ کاهش پیچیدگی محاسباتی	۲-۴-۴-۵	۲-۴-۴-۵ کاهش پیچیدگی محاسباتی
۶۲	..... ۳-۴-۴-۵ انعطاف پذیری در مقیاس های مختلف شبکه	۳-۴-۴-۵	۳-۴-۴-۵ انعطاف پذیری در مقیاس های مختلف شبکه
۶۳	..... ۴-۴-۴-۵ تطبیق پذیری با انواع شبکه ها	۴-۴-۴-۵	۴-۴-۴-۵ تطبیق پذیری با انواع شبکه ها

۶۳	۵-۴-۴-۵ زمان آموزش و اجرا بهینه
۶۳	۶-۴-۴-۵ قابلیت تفسیرپذیری
۶۳	۷-۴-۴-۵ عملکرد برتر نسبت به الگوریتم‌های پایه
۶۳	۸-۴-۴-۵ تطبیق با نیازهای کاربردی
۶۴	۵-۴-۴-۵ تحلیل مدت زمان آموزش و اجرا
۶۴	۱-۵-۴-۵ مدت زمان آموزش مدل‌ها
۶۵	۲-۵-۴-۵ مدت زمان تست و اعتبارسنجی مدل‌ها
۶۵	۳-۵-۴-۵ تحلیل کلی
۶۶	۶ جمع‌بندی و نتیجه‌گیری و پیشنهادات
۶۷	۱-۶ جمع‌بندی و نتیجه‌گیری
۶۸	۲-۶ پیشنهادات
۶۹	کتابنامه
۷۱	پیوست الف

صفحه	فهرست تصاویر	شکل
۲۱	۱-۳ نمونه ای از فرایند شبیه سازی	
۲۸	۲-۳ توزیع مقادیر برچسبها در گراف jazz	
	۳-۳ روند تغییرات تعداد گرههای مستعد ( $S$ ), آلوده ( $I$ ) و بهبود یافته ( $R$ ) در طول زمان	
۲۹	برای مقادیر مختلف $\beta$	
۲۹	۴-۳ تحلیل اهمیت گرهها در شبیهسازی SIR برای مقادیر مختلف $\beta$	
۳۱	۵-۳ اثر دقت مقادیر انتشار قدرت و توزیع آن در دقت‌های مختلف	
۳۲	۶-۳ تأثیر تعداد تکرار شبیهسازی بر روی قدرت انتشار ۸ گرههای تصادفی در گراف Figeys	
۳۳	۷-۳ زمان شبیهسازی گرافهای مختلف	
۳۴	۸-۳ تأثیر ترکیبی میانگین درجه ی گراف و تعداد گرههای گراف بر افزایش زمان شبیهسازی	
۳۵	۹-۳ ساختار تقسیم وظایف میان سرورها	
۳۶	۱۰-۳ نمایی از هموروش	
۴۰	۱-۴ مدل توسعه یافته	
۴۳	۲-۴ ویژگی‌های استخراج شده از گراف Figeys در فایل csv ذخیره شده	
۴۸	۳-۴ نمایش گرافها در رابط کاربری تعاملی برای تحلیل شبکه‌ها	
۴۸	۴-۴ رابط کاربری تعاملی برای تحلیل مقیاس تأثیرگذاری گرهها با استفاده از شبیهسازی SIR	
۵۷	۱-۵ توزیع مقادیر SIR در گرافهای مختلف (نمودار Boxplot و Min-Max)	
۵۸	۲-۵ عملکرد مدل در گرافهای دنیای واقعی بر اساس معیارهای مختلف	
۵۹	۳-۵ عملکرد مدل در گرافهای مصنوعی و بر اساس معیارهای مختلف	
۵۹	۴-۵ دقت مدل در معیارهای مختلف برای گرافهای مختلف در تنظیم test_L15_b4_sir2	
۶۱	۵-۵ مقایسه عملکرد مدل‌های پیشنهادی با الگوریتم‌های پایه در معیارهای مختلف.	
۶۴	۶-۵ مدت زمان آموزش مدل‌ها برای مقادیر مختلف پارامترهای L، sir_alpha، و epochs	
۶۵	۷-۵ مدت زمان تست مدل‌ها بر روی گرافهای مختلف	

## فهرست جداول

صفحه

جدول

۲۴	.....	۱-۳	جدول اطلاعات گراف های مصنوعی
۲۴	.....	۲-۳	جدول اطلاعات گراف های واقعی
۳۳	.....	۳-۳	نمونه خروجی استخراج ویژگی های حجمی
۳۳	.....	۴-۳	زمان محاسبه برای تولید برچسب برای یک گره در گراف 4_7000_4 ba_edgelist_exp1_7000_4
			به ازای تعداد تکرارهای مختلف.

# فصل اول

## مقدمه

## ۱-۱ مقدمه

## ۱-۱-۱ روند رو به رشد استفاده از الگوریتم‌های هوشمند در تحلیل شبکه‌های پیچیده

با پیشرفت سریع فناوری‌های اطلاعاتی و ارتباطی، بسیاری از حوزه‌ها به سمت استفاده از الگوریتم‌های هوشمند و یادگیری ماشین<sup>۱</sup> سوق پیدا کرده‌اند. تحلیل شبکه‌های پیچیده<sup>۲</sup> نیز یکی از زمینه‌هایی است که در سال‌های اخیر به شدت تحت تاثیر این پیشرفت‌ها قرار گرفته است. شبکه‌های پیچیده، که شامل گره‌ها<sup>۳</sup> و روابط پیچیده بین آن‌ها هستند، در بسیاری از مسائل از جمله تحلیل شبکه‌های اجتماعی، زیستی<sup>۴</sup> و اقتصادی به کار می‌روند. در این زمینه، استفاده از الگوریتم‌های هوشمند می‌تواند کمک کند تا الگوهای پیچیده و روابط پنهان در شبکه‌ها شناسایی شده و گره‌های تأثیرگذار<sup>۵</sup> رتبه‌بندی شوند.

استفاده از یادگیری ماشین و بهویژه شبکه‌های عصبی پیچشی<sup>۶</sup> در این تحلیل‌ها توانسته به طور قابل توجهی دقت و کارایی روش‌های تحلیل شبکه را بهبود بخشد. الگوریتم‌های شبکه‌های عصبی پیچشی با توانایی استخراج ویژگی‌های پیچیده و الگوهای پنهان از داده‌های شبکه، به شناسایی گره‌های کلیدی و رتبه‌بندی آن‌ها کمک می‌کنند.

## ۱-۱-۲ اهمیت شناسایی گره‌های تأثیرگذار

شناسایی گره‌های تأثیرگذار در شبکه‌های پیچیده از اهمیت بالایی برخوردار است (این گره‌ها معمولاً نقش مهمی در ساختار و عملکرد کلی شبکه ایفا می‌کنند و می‌توانند در فرایندهای تصمیم‌گیری، پیش‌بینی رفتار شبکه، و بهبود عملکرد شبکه‌های مختلف مورد استفاده قرار گیرند. بهویژه در شبکه‌های اجتماعی، شناسایی این گره‌ها می‌تواند به تحلیل تأثیرات فردی بر شبکه و پیش‌بینی الگوهای رفتار اجتماعی کمک کند).

در شبکه‌های زیستی نیز شناسایی گره‌های تأثیرگذار می‌تواند به کشف اهداف درمانی جدید و پیش‌بینی روند بیماری‌ها کمک کند. به همین دلیل، این یک حوزه مهم تحقیقاتی است که توجه بسیاری را به خود جلب کرده و مطالعات گسترده‌ای در آن صورت می‌گیرد.

<sup>1</sup>Machine Learning<sup>2</sup>Complex Networks<sup>3</sup>Node<sup>4</sup>Biological network<sup>5</sup>Influential nodes<sup>6</sup>Convolutional Neural Networks

### ۱-۳-۱ چالش‌های موجود در تحلیل شبکه‌های پیچیده

یکی از چالش‌های اصلی در تحلیل و پیاده‌سازی مدل‌ها برای شبکه‌های پیچیده، کمود داده‌های برچسب‌خورده<sup>۷</sup> با کیفیت بالا است که می‌تواند ارزیابی تأثیر گره‌ها را دشوار کند. این موضوع به‌ویژه در شبکه‌هایی با ساختارهای پیچیده مانند شبکه‌های چندلایه<sup>۸</sup> یا غیرهمگن<sup>۹</sup> به چشم می‌خورد، زیرا ایجاد مجموعه‌داده‌هایی که نماینده تمامی ویژگی‌های شبکه باشند، به منابع و تلاش زیادی نیاز دارد. علاوه بر این، پردازش این داده‌ها به نحوی که اطلاعات کلیدی و تأثیرگذار حفظ شود، چالشی دیگر است. استفاده از ویژگی‌هایی مانند شاخص اچ<sup>۱۰</sup> یا مرکزیت درجه<sup>۱۱</sup> نیازمند تنظیم دقیق پارامترها است و در شبکه‌های بزرگ با میلیون‌ها گره، به پیچیدگی محاسباتی بیشتری منجر می‌شود. شبکه‌های بزرگ و پیچیده همچنین اجرای مدل‌های مبتنی بر شبکه‌های عصبی پیچشی را به چالشی بزرگ تبدیل می‌کنند، زیرا این مدل‌ها نیازمند محاسبات سنگین ماتریسی برای تمامی گره‌ها هستند. تکنیک‌هایی مانند کاهش ابعاد<sup>۱۲</sup> یا فشرده‌سازی مدل<sup>۱۳</sup> می‌توانند به کاهش این چالش کمک کنند، اما ممکن است به از دست رفتن برخی از اطلاعات مهم منجر شوند. مدیریت منابع محاسباتی<sup>۱۴</sup>، بهینه‌سازی فرآیند یادگیری<sup>۱۵</sup> و استفاده از زیرساخت‌های قوی مانند پردازنده‌های گرافیکی<sup>۱۶</sup> و رایانش توزیع شده<sup>۱۷</sup> برای مقابله با این چالش‌ها ضروری است. همچنین، مدل‌سازی دقیق روابط گره‌ها و استخراج ویژگی‌های مناسب<sup>۱۸</sup> از آن‌ها اهمیت ویژه‌ای دارد، زیرا روابط در شبکه‌های پیچیده اغلب غیرخطی و دشوار برای شناسایی هستند. این امر نیاز به استفاده از روش‌های پیشرفته مانند شبکه‌های عصبی پیچشی را افزایش می‌دهد که قادر به شناسایی الگوهای پیچیده و روابط غیرخطی در شبکه‌ها هستند. در نهایت، اندازه‌گیری تأثیر هر گره نیز ممکن است با چالش‌های خاصی همراه باشد، زیرا تأثیر یک گره می‌تواند به عوامل مختلفی بستگی داشته و به طور مستقیم قابل اندازه‌گیری نباشد.

<sup>7</sup>Labeled Data

<sup>8</sup>Multilayer Networks

<sup>9</sup>Heterogeneous Networks

<sup>10</sup>H-Index Centrality

<sup>11</sup>Degree Centrality

<sup>12</sup>Dimensionality Reduction

<sup>13</sup>Model Compression

<sup>14</sup>Computational Resource Management

<sup>15</sup>Learning Process Optimization

<sup>16</sup>Graphics Processing Units (GPUs)

<sup>17</sup>Distributed Computing

<sup>18</sup>Feature Extraction

## ۴-۱ مزایای استفاده از شبکه‌های عصبی پیچشی

شبکه‌های عصبی پیچشی به ویژه برای تحلیل داده‌های تصویری<sup>۱۹</sup> و شبکه‌های گرافی<sup>۲۰</sup> بسیار مؤثر هستند. این شبکه‌ها قادرند ویژگی‌های پیچیده و پنهان داده‌ها را استخراج کرده و الگوهای تأثیرگذار را شناسایی کنند. در زمینه تحلیل شبکه‌های پیچیده، شبکه‌های عصبی پیچشی می‌توانند به شناسایی گره‌های تأثیرگذار و رتبه‌بندی آن‌ها بر اساس ویژگی‌های مختلف کمک کنند.

این الگوریتم‌ها با استفاده از لایه‌های پیچشی<sup>۲۱</sup> می‌توانند ویژگی‌های شبکه را به‌طور خودکار استخراج کنند، که باعث افزایش دقت و کارایی مدل‌های تحلیل شبکه می‌شود. به ویژه در شبکه‌های بزرگ و پیچیده، استفاده از این شبکه‌ها به‌طور قابل توجهی سرعت تحلیل و شناسایی گره‌های تأثیرگذار را افزایش می‌دهد.

## ۴-۱-۵ اهمیت پژوهش

با توجه به پیچیدگی و اهمیت تحلیل شبکه‌های تأثیرگذار می‌تواند تأثیرات قابل توجهی در بهبود عملکرد این شبکه‌ها داشته باشد. به ویژه در شبکه‌های اجتماعی و زیستی، شناسایی و رتبه‌بندی گره‌های تأثیرگذار می‌تواند به تصمیم‌گیری‌های بهتر و پیش‌بینی رفتارهای آینده کمک کند. از طرفی، استفاده از الگوریتم‌های هوشمند و به‌ویژه شبکه‌های عصبی پیچشی به‌طور چشمگیری دقت و کارایی این فرایندها را بهبود می‌بخشد. به همین دلیل، پژوهش در این زمینه می‌تواند به توسعه روش‌های دقیق‌تر و مؤثرتر برای تحلیل شبکه‌های پیچیده کمک کرده و به کاربردهای عملی در حوزه‌های مختلف مانند شبکه‌های اجتماعی، سیستم‌های زیستی و حتی سیستم‌های اقتصادی منجر شود.

## ۴-۱-۶ هدف پژوهش

هدف اصلی این پژوهه، توسعه یک چارچوب جدید است که بتواند گره‌های تأثیرگذار را در شبکه‌های پیچیده شناسایی و رتبه‌بندی کند. این چارچوب، مسئله شناسایی گره‌های تأثیرگذار را به عنوان یک مسئله کلاسیک رگرسیون<sup>۲۲</sup> مدل‌سازی کرده و از شبکه‌های عصبی پیچشی برای یادگیری ویژگی‌های ساختاری گره‌ها بهره می‌گیرد. به منظور بهبود دقت شناسایی و رتبه‌بندی گره‌ها، رویکردی چندمقیاسی<sup>۲۳</sup>

<sup>19</sup>Image Data

<sup>20</sup>Graph Networks

<sup>21</sup>Convolutional Layers

<sup>22</sup>Regression

<sup>23</sup>Multi-Scale Approach

برای استخراج و بازنمایی اطلاعات گره‌ها بر اساس همسایگان یک جهشی<sup>۲۴</sup> آن‌ها استفاده شده است. هدف این است که بازنمایی‌های محلی گره‌ها به طور کارآمدی ویژگی‌های ساختاری<sup>۲۵</sup> شبکه را منعکس کرده و اطلاعات بیشتری در مورد تأثیر گره‌ها ارائه دهد.

این چارچوب همچنین برای اطمینان از کارایی در مقیاس‌های بزرگ طراحی شده است. هدف این است که بتواند پس از آموزش روی شبکه‌های کوچک‌تر، عملکرد خود را به طور مؤثر به شبکه‌های واقعی<sup>۲۶</sup> و مصنوعی<sup>۲۷</sup> با اندازه‌ها و ویژگی‌های متنوع تعمیم دهد. برای تحقق این هدف، این پروژه از یک مدل استاندارد شبکه (مدل باراباشی–آلبرت<sup>۲۸</sup>) برای آموزش استفاده می‌کند و از معیارهای ساده‌ای مانند درجه و شاخص  $H$  برای ساخت مجموعه کانال‌های ساختاری بازنمایی گره‌ها<sup>۲۹</sup> بهره می‌گیرد. این اهداف به چارچوب اجازه می‌دهد تا رتبه‌بندی دقیقی از گره‌های تأثیرگذار ایجاد کند که هم در شبکه‌های کوچک و هم بزرگ مقیاس قابل استفاده باشد. علاوه بر این، به منظور افزایش سهولت استفاده و دسترسی کاربران، یک رابط کاربری<sup>۳۰</sup> برای این پروژه طراحی و پیاده‌سازی خواهد شد. این رابط کاربری به کاربران امکان می‌دهد تا به صورت بصری با چارچوب تعامل داشته باشند، شبکه‌های مختلف را بازگذاری و تحلیل کنند، گره‌های تأثیرگذار را مشاهده نمایند، و نتایج رتبه‌بندی را به صورت گرافیکی و قابل فهم دریافت کنند. هدف از توسعه این رابط کاربری، تسهیل فرآیند استفاده از چارچوب برای کاربران غیرمتخصص و فراهم کردن تجربه کاربری بهینه برای تحلیل شبکه‌های پیچیده است. این رابط کاربری نقش مهمی در کاربردی‌تر کردن چارچوب و گسترش استفاده از آن در زمینه‌های مختلف خواهد داشت.

## ۱-۷ ساختار پژوهش

ساختار این پروژه شامل چندین فصل است که از معرفی موضوع تا ارزیابی و نتیجه‌گیری را پوشش می‌دهد. ابتدا، مرور مختصری بر کارهای پیشین انجام می‌شود تا جایگاه این تحقیق مشخص شود. در ادامه، فصل مربوط به داده‌ها شامل قسمت‌های تعاریف، جمع‌آوری، پیش‌پردازش، تولید برچسب‌های گراف، و به طور کلی کار با داده‌ها است. فصل سامانه‌ی توسعه‌یافته به معرفی چارچوب و روش‌های پیشنهادی برای حل مسئله و جزئیات آن‌ها اختصاص دارد. نتایج حاصل از آزمایش‌ها و ارزیابی چارچوب در فصل ارزیابی ارائه

<sup>24</sup>One-hop neighborhood

<sup>25</sup>Topological features

<sup>26</sup>Real-World Networks

<sup>27</sup>Synthetic Networks

<sup>28</sup>Barabási–Albert model

<sup>29</sup>Structural channel sets of node representations

<sup>30</sup>User Interface (UI)

شده و در نهایت، جمع‌بندی، نتیجه‌گیری، و پیشنهادات برای کارهای آتی در فصل آخر بحث می‌شود.

## فصل دوم

### پیشینه موضوع

## ۱-۲ مروری بر تحقیقات پیشین و مطالعات مرتبط

در سال ۲۰۲۰، گوهنگ ژانو<sup>۱</sup>] تحقیقی با تمرکز بر شناسایی گره‌های تأثیرگذار در شبکه‌های پیچیده منتشر کردند. هدف این پژوهش، ارائه مدل یادگیری عمیق<sup>۱</sup> برای شناسایی گره‌های تأثیرگذار بود که ویژگی‌های ساختاری و ارتباطی شبکه را به صورت همزمان تحلیل کند. آن‌ها در این مطالعه از روش شبکه‌های پیچشی گراف<sup>۲</sup> بهره بردن. ویژگی‌های انتخابی شامل درجه، مرکزیت نزدیکی<sup>۳</sup>، مرکزیت بینابینی<sup>۴</sup>، و ضریب خوشبندی<sup>۵</sup> بودند. روش اعتبارسنجی، استفاده از شبیه‌سازی‌های مدل SIR بود. دادگان استفاده شده شامل پنج شبکه واقعی از انواع مختلف شبکه‌های اجتماعی، شبکه‌های پروتئینی، و شبکه‌های همکاری بود.

نتایج این مطالعه نشان داد که مدل ارائه شده<sup>۶</sup> توانست عملکرد پیش‌بینی را به طور قابل توجهی بهبود بخشد و گره‌های تأثیرگذار را با دقت بالاتری نسبت به روش‌های سنتی شناسایی کند. مزیت اصلی این پژوهش، ترکیب همزمان اطلاعات ساختاری و ویژگی‌های گره‌ها برای شناسایی دقیق‌تر گره‌های تأثیرگذار است. از طرفی دیگر، یکی از محدودیت‌های مقاله نیاز به داده‌های بزرگ<sup>۷</sup> و زمان پردازش بیشتر برای مدل‌سازی شبکه‌های کوچک بود. این تحقیق همچنین نشان داد که با پیش‌آموزش<sup>۸</sup> مدل بر روی شبکه‌های بزرگ و سپس تنظیم مجدد آن برای شبکه‌های کوچک، می‌توان عملکرد مدل را در شبکه‌های با داده‌های کمتر بهبود بخشید.

در سال ۲۰۲۲، یانگ او و همکارانش<sup>۹</sup>] تحقیقی با تمرکز بر شناسایی گره‌های تأثیرگذار در شبکه‌های پیچیده منتشر کردند. هدف این تحقیق توسعه الگوریتمی کارآمد و دقیق برای شناسایی گره‌های تأثیرگذار در فرایند انتشار اطلاعات بود. آن‌ها از الگوریتم شبکه عصبی پیچشی گرافی چندکاناله<sup>۹</sup> استفاده کردند که شامل ویژگی ساختاری در سه سطح میکرو<sup>۱۰</sup>، جامعه<sup>۱۱</sup>، و ماکرو<sup>۱۲</sup> بود. ویژگی‌های انتخابی

<sup>1</sup>Deep Learning Model

<sup>2</sup>Graph Convolutional Networks

<sup>3</sup>Closeness Centrality

<sup>4</sup>Betweenness Centrality

<sup>5</sup>Clustering Coefficient

<sup>6</sup>InfGCN

<sup>7</sup>Big Data

<sup>8</sup>Pre-Training

<sup>9</sup>Multi-channel Recurrent Convolutional Neural Network

<sup>10</sup>Micro Structural Features

<sup>11</sup>Global Structural Features

<sup>12</sup>Macro Structural Features

شامل درجه همسایگان، تعداد جوامع متصل به گره و مرکزیت هسته ای<sup>۱۳</sup> بود. برای اعتبارسنجی این پژوهش، از مدل SIR برای تولید برچسبها استفاده شد و داده‌های استفاده شده شامل ۹ شبکه واقعی و شبکه‌های مصنوعی باراباشی-آلبرت بود.

نتایج این تحقیق نشان داد که الگوریتم M-RCNN به طور میانگین دقت شناسایی<sup>۱۴</sup> را ۹۷/۲۵٪ نسبت به الگوریتم RCNN بهبود بخشدید و در عین حال پیچیدگی محاسباتی<sup>۱۵</sup> مشابهی داشت. مزیت اصلی این تحقیق توانایی ترکیب خودکار اطلاعات ساختاری چندسطحی بدون نیاز به وزن دهی از پیش تعریف شده است. از طرفی، عملکرد این الگوریتم ممکن است در شبکه‌های پراکنده<sup>۱۶</sup> یا با ساختارهای بسیار پیچیده دچار کاهش شود. نتایج این مطالعه درک عمیق‌تری از تأثیر ساختار شبکه بر عملکرد شبکه عصبی پیچشی گرافی ارائه می‌دهد و کاربردهایی در شناسایی گره‌های کلیدی در انواع شبکه‌ها دارد.

در سال ۲۰۲۲، آن-یو یو، یان فو و دوآن-بینگ چن<sup>۱۷</sup> تحقیقی با تمرکز بر شناسایی گره‌های کلیدی<sup>۱۸</sup> در شبکه‌های پیچیده با استفاده از یادگیری گراف<sup>۱۹</sup> منتشر کردند. هدف این پژوهش کاهش اندازه مجموعه اولیه گره‌های انتشار<sup>۲۰</sup> برای دستیابی به میزان آلدگی مشخص<sup>۲۱</sup> در شبکه‌های پیچیده بود. در این مطالعه، از چارچوب یادگیری عمیق گراف به نام IMGN<sup>۲۲</sup> استفاده شد. ویژگی‌های انتخابی شامل مرکزیت گره‌ها مانند مرکزیت درجه، مرکزیت رتبه‌بندی صفحه ای<sup>۲۳</sup> و مرکزیت هسته‌ای<sup>۲۴</sup> بود. برای اعتبارسنجی مدل از روش‌های شبیه‌سازی بر پایه مدل انتشار SIR استفاده شد و دادگان مورد استفاده شامل شبکه‌های مصنوعی<sup>۲۵</sup> و پنج شبکه واقعی<sup>۲۶</sup> از جمله شبکه‌های اجتماعی و شبکه‌های حمل و نقل بود.

نتایج این پژوهش نشان داد که IMGN در مقایسه با الگوریتم‌های سنتی رتبه‌بندی گره‌های

<sup>۱۳</sup>k-core centrality

<sup>۱۴</sup>Identification Accuracy

<sup>۱۵</sup>Computational Complexity

<sup>۱۶</sup>Sparse Networks

<sup>۱۷</sup>Critical Nodes

<sup>۱۸</sup>Graph Learning

<sup>۱۹</sup>initial propagation set

<sup>۲۰</sup>fixed infection scale

<sup>۲۱</sup>Information Maximization with Graph Neural Networks

<sup>۲۲</sup>PageRank

<sup>۲۳</sup>k-core decomposition

<sup>۲۴</sup>Synthetic Network

<sup>۲۵</sup>real-world networks

غیرتکراری<sup>۲۶</sup> مانند تبہبندی صفحه ای و هسته‌بندی  $k^{27}$ ، کمترین نسبت گره‌های اولیه را برای دستیابی به میزان آلودگی بیش از  $80^{\circ}$  درصد دارد. این نکته، نشان‌دهنده‌ی کارایی روش در بهینه‌سازی فرآیند انتشار است. مزیت اصلی این پژوهش توانایی یادگیری ویژگی‌های کلیدی گره‌ها از شبکه‌های کوچک مصنوعی و تعمیم آن به شبکه‌های بزرگ‌تر بود. از سوی دیگر، پیچیدگی محاسباتی بالا برای تولید مجموعه داده‌های آموزشی محدودیت این روش است. علاوه بر این، استفاده از الگوریتم بازترتبی<sup>۲۸</sup> RINF برای بازترتبی نتایج IMGNN نشان داد که این چارچوب می‌تواند با بهبود عملکرد خود به‌طور موثری در شناسایی گره‌های کلیدی عمل کند.

در سال ۲۰۲۲، لیو و کائو<sup>۲۹</sup>] تحقیقی با تمرکز بر رتبه‌بندی گره‌ها در شبکه‌های پیچیده منتشر کردند. هدف این تحقیق توسعه یک مدل یادگیری به رتبه‌بندی بر اساس شبکه‌های پیچیده بود که بتواند اهمیت گره‌ها را با استفاده از اطلاعات ساختاری شبکه و ویژگی‌های گره‌ها تعیین کند. آن‌ها در این مطالعه از مدل‌های یادگیری خودنظراتی<sup>۳۰</sup> و مدل پیچشی گراف استفاده کردند. ویژگی‌های انتخابی شامل اطلاعات محلی گره‌ها و ساختار شبکه بود. روش اعتبارسنجی، آزمایش‌های گسترده بر روی مجموعه داده‌های واقعی بود. دادگان مورد استفاده شامل شبکه‌های همکاری، شبکه ویکی‌پدیا، و داده‌های مترو بود.

نتایج این مطالعه نشان داد که مدل ارائه شده در مقایسه با روش‌های سنتی و مدل‌های یادگیری عمیق دیگر عملکرد بهتری دارد و پایداری بیشتری در مجموعه داده‌های مختلف از خود نشان می‌دهد. مزیت اصلی این پژوهش استفاده از یادگیری وظایف چندگانه<sup>۳۱</sup> و یادگیری خودنظراتی برای بهبود دقت رتبه‌بندی گره‌ها بود. از طرفی دیگر، وابستگی به داده‌های برچسب‌گذاری شده یکی از محدودیت‌های این مقاله بود. این پژوهش نه تنها اهمیت گره‌ها را بر اساس اطلاعات شبکه تحلیل می‌کند، بلکه قادر است با استفاده از تعداد کمی داده برچسب‌گذاری شده نتایج قابل اعتمادی ارائه دهد.

در سال ۲۰۲۴، واسیم احمد و بانگ وانگ<sup>۳۲</sup>] تحقیقی با تمرکز بر شناسایی گره‌های مؤثر در شبکه‌های پیچیده منتشر کردند. هدف این پژوهش بهبود فرآیند انتشار اطلاعات از طریق شناسایی گره‌هایی بود که تأثیر زیادی در ساختار شبکه دارند. آن‌ها در این مطالعه از یک چارچوب جدید به نام LCN<sup>۳۳</sup> استفاده کردند که مبتنی بر شبکه‌های عصبی پیچشی و نمایش‌های محلی گره‌ها است. روش اعتبارسنجی این

<sup>26</sup>non-iterative node ranking algorithms

<sup>27</sup>Kshell

<sup>28</sup>Reordering Algorithm

<sup>29</sup>Self-supervised learning

<sup>30</sup>Multi-task learning

پژوهش نیز مانند پژوهش های قبلی، شبیه‌سازی مدل SIR بود و داده‌گان شامل شبکه‌های واقعی و مصنوعی مختلف بودند.

نتایج این مطالعه نشان داد که چارچوب LCNN نسبت به روش‌های پیشرفته دیگر عملکرد بهتری در شناسایی گره‌های مؤثر دارد و زمان اجرای متوسط آن نیز مناسب شبکه‌های بزرگ‌مقیاس است. مزیت اصلی این پژوهش، کاهش زمان محاسبات و بهبود نمایش محلی گره‌ها بود. از طرفی دیگر، محدودیت این مقاله استفاده از ویژگی‌های محلی بود که ممکن است برخی از جنبه‌های جامع شبکه را نادیده بگیرد. در نهایت، این چارچوب به دلیل استفاده از شبکه‌های مصنوعی برای آموزش و قابلیت تعمیم به شبکه‌های واقعی، کاربردهای متعددی در شبکه‌های اجتماعی و بازاریابی ویروسی‌وار<sup>۳۱</sup> دارد.

---

<sup>31</sup>Viral Marketing

## فصل سوم

### داده‌ها

## ۱-۳ ساختار فصل

در این فصل، ابتدا مفاهیم پایه‌ای که برای مدیریت و پردازش داده‌ها در این پژوهش مورد استفاده قرار گرفته‌اند را توضیح می‌دهیم.

سپس وارد بخش کار با داده می‌شویم، جایی که مراحل آماده‌سازی و پردازش داده‌ها به تفصیل شرح داده می‌شوند. این مراحل شامل انتخاب و ساخت مجموعه داده‌ها، تولید گراف‌های باراباشی-آلبرت، پیش‌پردازش داده‌ها<sup>۱</sup>، استخراج ویژگی‌های گراف، و ساخت برچسب‌های SIR است. در این بخش همچنین به روش‌های بهینه‌سازی فرآیند تولید برچسب‌ها و بهبود سرعت پردازش با استفاده از ابزارهای موازی‌سازی<sup>۲</sup> پرداخته خواهد شد.

## ۲-۳ تعاریف

در این بخش، مفاهیم و ابزارهای اصلی که در این پژوهش برای مدیریت و پردازش داده‌ها به کار رفته‌اند، معرفی و توضیح داده می‌شوند. ابتدا مدل گراف باراباسی-آلبرت به عنوان یک مدل گراف بدون مقیاس<sup>۳</sup> معرفی می‌شود. سپس، مدل شبیه‌سازی SIR شرح داده می‌شود که برای برچسب‌گذاری گره‌های گراف و مدل‌سازی انتشار اطلاعات یا بیماری در شبکه‌ها استفاده شده است. در نهایت، داکر<sup>۴</sup> به عنوان یک ابزار کاربردی برای ایجاد و مدیریت محیط‌های مجازی مورد بررسی قرار می‌گیرد.

## ۱-۲-۳ مدل گراف باراباسی-آلبرت

مدل باراباسی-آلبرت<sup>[۶]</sup> (یا به اختصار BA) یک مدل گراف بدون مقیاس است که برای توصیف شبکه‌هایی با خاصیت توزیع توانی<sup>۵</sup> استفاده می‌شود. این مدل توسط آلبرت لازلو باراباسی<sup>۶</sup> و رکا آلبرت<sup>۷</sup> در سال ۱۹۹۹ معرفی شد و یکی از محبوب‌ترین مدل‌ها برای مطالعه شبکه‌های پیچیده است. این مدل نشان می‌دهد که در بسیاری از شبکه‌های واقعی مانند اینترنت، شبکه‌های اجتماعی، و شبکه‌های زیستی، تعداد زیادی از گره‌ها دارای تعداد کمی اتصال هستند، در حالی که تعداد کمی از گره‌ها (که به

<sup>1</sup>Data Preprocessing

<sup>2</sup>Parallelization

<sup>3</sup>Scale-Free Network

<sup>4</sup>Docker

<sup>5</sup>Power-Law Distribution

<sup>6</sup>Albert-László Barabási

<sup>7</sup>Réka Albert

آن‌ها «ابرگره‌ها»<sup>۸</sup> گفته می‌شود) دارای تعداد زیادی اتصال هستند. این خاصیت به گراف ویژگی «بدون مقیاس» می‌دهد، به این معنا که توزیع درجات گره‌ها فاقد مقیاس مشخص است و به صورت توزیع توان<sup>۹</sup> با احتمال  $P(k) \sim k^{-\gamma}$  توصیف می‌شود، که درجه گره و نمای توانی<sup>۱۰</sup>  $\gamma$  معمولاً بین ۲ و ۳ است.

### ۱. رشد<sup>۱۱</sup>

- گراف با تعداد کمی از گره‌ها ( $m_0$ ) شروع می‌شود که به صورت کامل<sup>۱۲</sup> به یکدیگر متصل هستند.

- در هر گام، یک گره جدید به گراف اضافه می‌شود که به  $m$  گره موجود در گراف متصل می‌شود ( $m \leq m_0$ ).

### ۲. ترجیح اتصال<sup>۱۳</sup>

- گره جدید ترجیح می‌دهد به گره‌هایی متصل شود که از قبل دارای تعداد بیشتری اتصال هستند.

- احتمال اتصال یک گره جدید ( $i$ ) به یک گره موجود ( $j$ ) متناسب با درجه گره  $j$  است:

$$\Pi(k_j) = \frac{k_j}{\sum k_i}$$

- در اینجا  $k_j$  تعداد یال<sup>۱۴</sup>‌های گره  $j$  و  $\sum k_i$  مجموع درجات تمام گره‌ها در گراف است.

- این ویژگی باعث ایجاد ابرگره‌هایی می‌شود که مرکز شبکه را تشکیل می‌دهند.

### ۱-۱-۲-۳ ویژگی‌های اصلی

- توزیع توانی: درجات گره‌ها در این مدل از توزیع زیر پیروی می‌کنند:

$$P(k) \sim k^{-\gamma}$$

<sup>8</sup>Hubs

<sup>9</sup>Power-Law Distribution

<sup>10</sup>Power-Law Exponent

<sup>11</sup>Growth

<sup>12</sup>Fully Connected

<sup>13</sup>Preferential Attachment

<sup>14</sup>Edge

این بدان معناست که احتمال یافتن گره‌ای با درجه بسیار بالا کوچک است، اما صفر نیست، و ابرگره‌ها نقش مهمی در ساختار شبکه ایفا می‌کنند.

- شبهات به شبکه‌های واقعی: بسیاری از شبکه‌های واقعی مانند وب، شبکه‌های اجتماعی، و شبکه‌های زیستی از توزیع توان-قانونی تبعیت می‌کنند و مدل باراباشی-آلبرت می‌تواند این ساختارها را شبیه‌سازی کند.

• اثر اتصال اولیه<sup>۱۵</sup>: گره‌هایی که در مراحل اولیه رشد گراف اضافه شده‌اند، تمایل بیشتری به تبدیل شدن به ابرگره دارند، زیرا فرصت بیشتری برای جذب یال‌های جدید داشته‌اند.

• کارایی بالا: با وجود پیچیدگی شبکه، گراف‌های BA دارای مسیرهای کوتاه میانگین<sup>۱۶</sup> و ضریب خوشبندی پایین هستند، که شبهات زیادی به شبکه‌های واقعی دارد.

### ۲-۱-۲-۳ مزايا و محدوديتها

این مدل مزايا و محدوديت هاي دارد که در قسمت زير ذكر شده اند.

#### • مزايا:

۱. سادگي و شهرت: مدل باراباشی-آلبرت ساده است و می‌تواند بسیاری از ویژگی‌های شبکه‌های واقعی را بازسازی کند.

۲. شبهات به واقعیت: ویژگی بدون مقیاس، آن را برای تحلیل شبکه‌هایی مانند شبکه‌های اجتماعی، ارتباطات اینترنتی و زیستی مناسب می‌کند.

#### • محدوديتها:

۱. فقدان خوشبندی<sup>۱۷</sup>: در گراف‌های BA، خوشبندی پایین‌تر از آن چیزی است که در بسیاری از شبکه‌های واقعی مشاهده می‌شود.

۲. فرض خطی بودن ترجیح اتصال: مدل فرض می‌کند احتمال اتصال مستقیماً متناسب با درجه گره است، اما در برخی شبکه‌ها، این رابطه غیرخطی است.

<sup>15</sup>Initial Connectivity Effect

<sup>16</sup>Short Average Path Length

<sup>17</sup>Clustering

۳. عدم توجه به وزن یال‌ها: مدل یال‌ها را بدون وزن در نظر می‌گیرد، در حالی که در بسیاری از شبکه‌های واقعی یال‌ها وزن دار هستند.

### ۲-۲-۳ مدل شبیه‌سازی SIR

شبیه‌سازی SIR (مخفف Susceptible-Infected-Recovered یا مستعد - آلوده - بهبود یافته) یکی از مدل‌های اصلی در تحلیل گسترش بیماری‌ها یا اطلاعات در شبکه‌های پیچیده است. این مدل که ابتدا برای مطالعه گسترش بیماری‌های عفونی طراحی شده بود، به دلیل انعطاف‌پذیری بالا و کاربرد گسترده‌اش در حوزه‌های مختلف، به یکی از ابزارهای اساسی در تحلیل‌های شبکه تبدیل شده است. در پژوهه حاضر، از این شبیه‌سازی برای ایجاد برچسب‌های مربوط به میزان اهمیت گره‌ها در شبکه‌های پیچیده استفاده شده است. این بخش توضیحات کاملی در مورد این نوع شبیه‌سازی، ویژگی‌ها، مزايا، محدودیت‌ها و دلیل استفاده از آن برای برچسب‌گذاری اهمیت (قدرت انتشار<sup>۱۸</sup>) گره‌ها ارائه می‌دهد.

### ۱-۲-۳ اجزای مدل

در این مدل پخش بیماری در شبکه، سه حالت برای هر گره در شبکه تعریف شده است:

- مستعد (S یا Susceptible): گره‌هایی که مستعد ابتلا به عفونت هستند
- آلوده (I یا Infected): گره‌هایی که در حال حاضر آلوده‌اند و می‌توانند عفونت را به گره‌های مستعد منتقل کنند
- بهبود یافته (R یا Recovered): گره‌هایی که پس از ابتلا به عفونت بهبود یافته و دیگر توانایی انتقال یا ابتلا به بیماری را ندارند

بدین ترتیب دستگاه معادلات در این مدل به شکل زیر درمی‌آیند:

- تغییر تعداد مستعدها:

$$\frac{dS}{dt} = -\beta SI$$

تعداد مستعدها با افزایش افراد آلوده کاهش می‌یابد و نرخ کاهش وابسته به تعداد مستعدها (S)، تعداد آلوده‌ها (I)، و  $\beta$  نرخ آلودگی (یا انتقال)<sup>۱۹</sup> است.

<sup>18</sup>Influential Scale

<sup>19</sup>Infection Rate

- تغییر تعداد آلودها:

$$\frac{dI}{dt} = \beta SI - \gamma I$$

تعداد آلودها با آلوده شدن مستعدها افزایش می‌یابد و با یهیوید آلودها کاهش می‌یابد. افزایش وابسته به نرخ آلودگی ( $\beta$ ) و کاهش وابسته به نرخ بهبودی ( $\gamma$ )<sup>۲۰</sup> است.

- تغییر تعداد بهبود یافته‌ها:

$$\frac{dR}{dt} = \gamma I$$

تعداد بهبود یافته‌ها با نرخ بهبودی ( $\gamma$ ) و تعداد آلودها (I) افزایش می‌یابد.

## ۲-۲-۲-۳ آستانه انتشار و نظریه میدان میانگین

نظریه میدان میانگین: هر گره در شبکه تحت تأثیر میانگین رفتار تمام گره‌های دیگر قرار دارد. به این ترتیب، نرخ انتشار بیماری فقط به ویژگی‌های کلی شبکه وابسته است. برای تعیین اینکه آیا بیماری یا اطلاعات به طور گستردگی منتشر خواهد شد، از مقدار آستانه انتشار  $\beta_c$ <sup>۲۱</sup> استفاده می‌شود.

$$\beta_c = \frac{\langle k \rangle}{\langle k^2 \rangle - \langle k \rangle}$$

در این معادله:

میانگین درجه گره‌ها :  $\langle k \rangle$

میانگین مربع درجه گره‌ها :  $\langle k^2 \rangle$

اگر  $\beta_c > \beta$ ، انتشار در شبکه می‌تواند فراگیر شود. برای درک این موضوع باید تعریف آستانه انتشار را بررسی کنیم. آستانه انتشار ( $\beta_c$ ) به مقداری از نرخ انتقال ( $\beta$ ) اطلاق می‌شود که در آن، یک بیماری توانایی گسترش پایدار در شبکه را به دست می‌آورد. این مقدار، حداقل نرخ انتقالی است که برای بقای SIR، بیماری و گسترش آن به کل شبکه مورد نیاز است. در مدل‌های شبیه‌سازی انتشار بیماری مانند  $\beta_c$  مشخص می‌کند که بیماری از چند فرد محدود به کل شبکه سرایت می‌کند یا به سرعت فروکش می‌کند<sup>۲۲</sup>. معادله‌ی آستانه انتشار نشان می‌دهد که هرچه میانگین مربع درجه گره‌ها بزرگ‌تر باشد

<sup>20</sup>Recovery Rate

<sup>21</sup>Infection Probability Threshold

(که در شبکه‌های مقیاس آزاد رایج است)، مقدار  $\beta_c$  کاهش می‌یابد و بیماری راحت‌تر در شبکه گسترش می‌یابد. آستانه انتشار سه ویژگی کلیدی را درباره رفتار شبکه و فرآیند انتشار مشخص می‌کند [۷، ۸]:

۱. ساختار شبکه: مقدار  $\beta_c$  به ساختار شبکه و توزیع درجات گره‌ها بستگی دارد. به عنوان مثال، در شبکه‌های مقیاس آزاد، که درجات گره‌ها توزیع بسیار نامتقارن دارند، مقدار  $\beta_c$  به شدت پایین است، زیرا گره‌های با درجه بالا (ابرگره‌ها) انتشار را تسهیل می‌کنند.
۲. پایداری انتشار: وقتی  $\beta_c < \beta$ ، انتشار می‌تواند به طور پایدار در شبکه ادامه یابد. این مقدار به عنوان یک مرز بین دو فاز «فروکش بیماری» و «گسترش بیماری» عمل می‌کند.
۳. اثر دینامیکی <sup>۲۲</sup>: مقدار  $\beta_c$  تأثیر نرخ انتقال بر رفتار شبکه را در طول زمان روشن می‌سازد، به ویژه در شبکه‌های بزرگ یا جهان‌کوچک <sup>۲۳</sup> که رفتار دینامیک شبکه بسیار وابسته به  $\beta$  است.

حال که با آستانه انتشار و ویژگی‌های آن آشنا هستیم، به تحلیل بازه‌ی ویژه‌ی  $1/\alpha \leq \beta \leq 1$  برای می‌پردازیم:

$$\beta = \alpha \cdot \beta_c \quad \text{که در آن} \quad \alpha \in [1, 1/\alpha]$$

- بررسی رفتار در نزدیکی آستانه: در این بازه، مقادیر  $\beta$  نزدیک به مقدار بحرانی  $\beta_c$  قرار دارند. این ناحیه بسیار حساس است و به ما اجازه می‌دهد رفتار شبکه در نزدیکترین حالت به تغییر فاز دینامیکی (از فروکش به انتشار پایدار) را مطالعه کنیم.

- بررسی انتقال فاز: مقادیر  $\beta$  بزرگ‌تر از  $\beta_c$  نشان‌دهنده حرکت از حالت فروکش به انتشار است. این بازه، رفتار شبکه را در هنگام عبور از این تغییر فاز بررسی می‌کند.

- اهمیت ساختار شبکه: همانطور که در تحقیقات پیشین [۷، ۸] ذکر شده، توزیع درجات گره‌ها و توپولوژی (ساختار) شبکه می‌تواند مقادیر مختلفی از  $\beta$  را برای پایداری انتشار ایجاد کند. بنابراین، مقادیر نزدیک و بالاتر از  $\beta_c$  به ما دیدگاه دقیق‌تری از اثر توپولوژی می‌دهد.

### ۳-۲-۳ شبیه‌سازی

فرایند شبیه‌سازی بر اساس انتقال بین این حالات تعریف می‌شود. در هر مرحله از شبیه‌سازی:

<sup>22</sup>Dynamic Effect

<sup>23</sup>Small-World Network

- گره‌های آلوده با احتمال  $\beta$  (نرخ انتقال) می‌توانند گره‌های مستعد همسایه را آلوده کنند.

- گره‌های آلوده با احتمال  $\gamma$  (نرخ بهبودی) به حالت بهبود یافته منتقل می‌شوند.

#### ۴-۲-۳ هدف از شبیه‌سازی

هدف از اجرای شبیه‌سازی SIR در این پروژه، ارزیابی نقش گره‌های مختلف در گسترش بیماری یا اطلاعات است. در واقع، اهمیت گره‌ها بر اساس تاثیر آن‌ها در گسترش عفونت سنجیده می‌شود. برای هر گره برجسب قدرت انتشار به نسبت تعداد گره‌هایی که از طریق آن آلوده (یا بهبود یافته) می‌شوند محاسبه می‌شود. این رویکرد به شناسایی گره‌های کلیدی کمک می‌کند که نقش مهمی در گسترش اطلاعات یا بیماری دارند.

#### ۴-۲-۵ فرمول تأثیرگذاری گره‌ها در مدل SIR

برای تعیین قدرت انتشار (تأثیرگذاری) یک گره (IS) در انتشار اطلاعات، از فرمول زیر استفاده می‌شود:

$$IS = \frac{1}{M} \sum_{i=1}^M \frac{\eta_R}{n} \quad (1-3)$$

که در آن:

- $M$ : تعداد شبیه‌سازی‌های انجام شده.

- $\eta_R$ : تعداد گره‌هایی که در پایان فرایند بهبود یافته‌اند.

- $n$ : تعداد کل گره‌های شبکه.

#### ۴-۲-۶ مزایا و محدودیت‌ها

- مزایا:

1. سادگی و انعطاف‌پذیری: مدل SIR ساختاری ساده دارد و می‌تواند به آسانی برای انواع مختلف شبکه‌ها و مسائل تنظیم شود.

2. تحلیل کمی: این مدل امکان اندازه‌گیری تاثیر گره‌ها را به صورت عددی فراهم می‌کند.

۳. کاربرد گستردده: از گسترش بیماری‌ها گرفته تا انتشار اطلاعات و شایعات، این مدل قابل استفاده است.

• محدودیت‌ها:

۱. پارامترهای ثابت: پارامترهای نرخ انتقال  $\beta$  و نرخ بهبود  $\gamma$  در طول شبیه‌سازی ثابت هستند و تغییرات احتمالی آن‌ها در دنیای واقعی را در نظر نمی‌گیرند.

۲. عدم در نظر گرفتن ساختار زمانی: این مدل روابط پیچیده‌تر مانند تاخیر در انتقال بیماری را لحاظ نمی‌کند.

۳. سادگی بیش از حد: این مدل عوامل خارجی مانند ایمنی گره‌ها یا تعاملات بین شبکه‌ها را در نظر نمی‌گیرد.

### ۷-۲-۲-۳ فرآیند شبیه‌سازی

در زمان  $t = 0$ ، یک گره به عنوان منبع اولیه آلوده انتخاب می‌شود.

در هر دوره<sup>۵</sup>:

- گره‌های آلوده همسایگان مستعد خود را با احتمال  $\beta$  آلوده می‌کنند.

- گره‌های آلوده با احتمال  $\gamma$  بهبود می‌یابند.

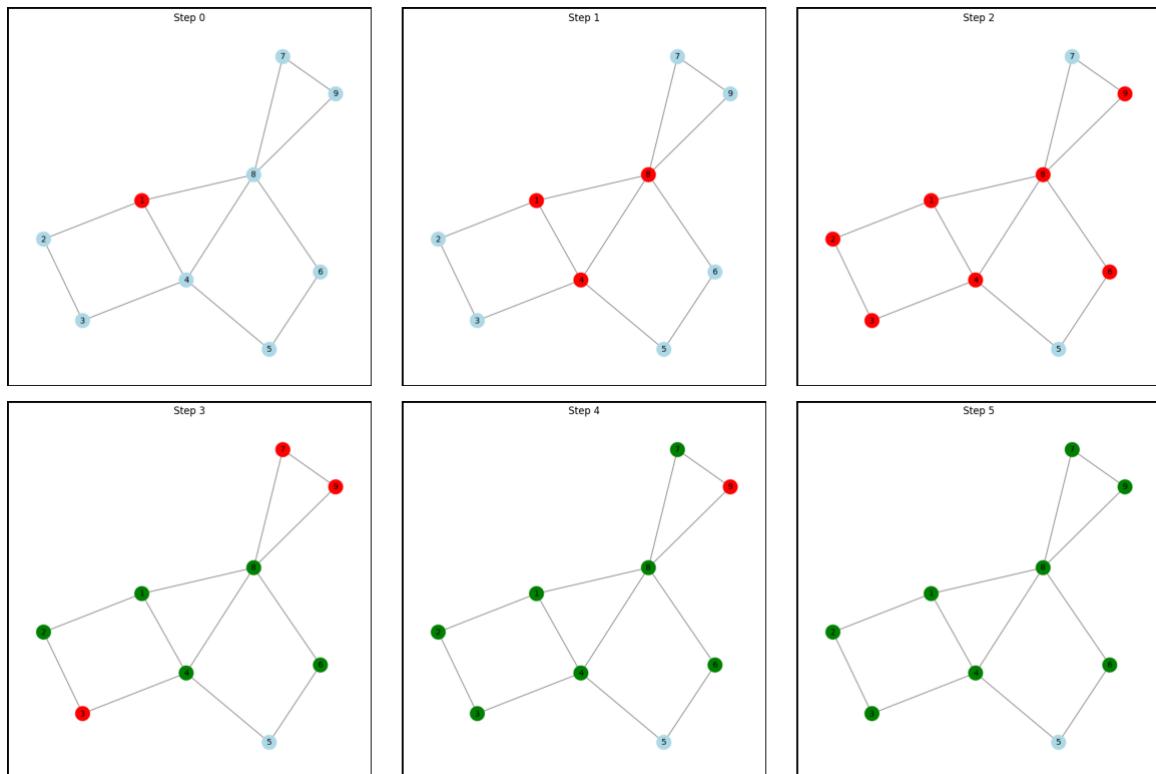
فرآیند ادامه می‌یابد تا زمانی که دیگر هیچ گره آلوده‌ای باقی نماند یا مقدار زمان مشخصی گذشته باشد. در ادامه در شکل ۱-۳ یک نمونه از این نوع شبیه‌سازی را مشاهده می‌کنیم.

### ۳-۲-۳ داکر

داکر یک پلتفرم متن‌باز<sup>۲۴</sup> است که به توسعه‌دهندگان و تیم‌های عملیاتی اجازه می‌دهد نرم‌افزارها را در قالب بسته‌های کوچک و قابل حمل به نام کانتینر<sup>۲۵</sup> اجرا و مدیریت کنند. کانتینرها شامل تمامی وابستگی‌ها و کتابخانه‌های لازم برای اجرای یک برنامه هستند و این اطمینان را فراهم می‌کنند که

<sup>24</sup>Open-Source

<sup>25</sup>Containers



شکل ۱-۳: نمونه‌ای از فرایند شبیه سازی

نرم‌افزار بدون توجه به سیستم‌عامل<sup>۲۶</sup> یا محیط اجراء<sup>۲۷</sup> به درستی کار می‌کند. این قابلیت به خصوص در سناریوهایی که محیط‌های مختلفی مانند توسعه<sup>۲۸</sup>، تست<sup>۲۹</sup> و تولید<sup>۳۰</sup> وجود دارد، بسیار مفید است.

### ۱-۳-۲-۳ ویژگی‌های کلیدی

- مقیاس‌پذیری<sup>۳۱</sup> و پایداری<sup>۳۲</sup>: یکی از مهم‌ترین ویژگی‌های داکر، مقیاس‌پذیری آن است. کانتینرهای داکر می‌توانند به راحتی در سرورهای مختلف مقیاس‌پذیر شوند و از یک سیستم به سیستم دیگر منتقل شوند، بدون اینکه مشکلی در سازگاری<sup>۳۳</sup> به وجود بیاید. این قابلیت باعث می‌شود که تیم‌ها بتوانند با سرعت بیشتری به تغییرات نیازهای پروژه پاسخ دهنند.

<sup>26</sup>Operating System (OS)

<sup>27</sup>Execution Environment

<sup>28</sup>Development Environment

<sup>29</sup>Testing Environment

<sup>30</sup>Production Environment

<sup>31</sup>Scalability

<sup>32</sup>Stability

<sup>33</sup>Compatibility

- مدیریت وابستگی‌ها<sup>۳۴</sup>: داکر تمامی وابستگی‌های لازم برای اجرای یک نرمافزار را در کانتینرها قرار می‌دهد. این ویژگی باعث می‌شود که توسعه‌دهندگان نگران مشکلاتی مانند ناسازگاری نسخه‌ها یا کمبود بسته‌های مورد نیاز نباشند. کانتینرها همچنین قابل حمل<sup>۳۵</sup> هستند، به این معنی که می‌توان آن‌ها را روی هر سیستمی که از داکر پشتیبانی می‌کند، اجرا کرد.
- یکپارچگی<sup>۳۶</sup> و تست آسان‌تر: داکر به تیم‌های توسعه اجازه می‌دهد تا یک محیط یکپارچه و استاندارد برای توسعه و تست نرمافزارها ایجاد کنند. این به معنی این است که کدها در تمامی مراحل توسعه و استقرار<sup>۳۷</sup> به طور یکسان عمل خواهند کرد، که به کاهش خطاهای افزایش کارایی کمک می‌کند.

### ۳-۳ مجموعه‌ی داده‌ها

#### ۱-۳-۳ کار با داده

برای پیاده‌سازی این پروژه، داده‌هایی شامل گراف‌های مصنوعی با رابطه‌ی آلبرت و گراف‌های دنیای واقعی، شبیه‌سازی‌های مدل SIR، و برچسب‌های تأثیرگذاری گره‌ها مورد نیاز بود. گراف‌های مصنوعی با استفاده از کتابخانه NetworkX تولید شدند، که امکان تنظیم پارامترهایی نظری تعداد گره‌ها و میانگین درجه را فراهم می‌کند. علاوه بر این، مدل SIR برای شبیه‌سازی انتشار و تعیین تأثیرگذاری گره‌ها در شبکه استفاده شد. داده‌های تولید شده از طریق این روش‌ها به عنوان ورودی برای مراحل بعدی پروژه، شامل استخراج ویژگی‌ها و ارزیابی عملکرد مدل، به کار گرفته شدند. در این بخش این مجموعه داده‌ها را بیشتر بررسی کرده‌ایم.

#### ۱-۱-۳ انتخاب، ساخت یا جمع آوری مجموعه‌ی داده‌ها

برای این پروژه، از دو نوع گراف استفاده شد: گراف‌های دنیای واقعی و گراف‌های مصنوعی (BA). گراف‌های دنیای واقعی با دقت و بر اساس گراف‌هایی انتخاب شدند که در مقالات معتبر<sup>[۱، ۲، ۳، ۴، ۵]</sup> مرتبط با این حوزه استفاده شده بودند. این انتخاب به دو دلیل اساسی صورت گرفت: نخست، امکان

<sup>34</sup>Dependency management

<sup>35</sup>Portable

<sup>36</sup>Integration

<sup>37</sup>Deployment

مقایسه مستقیم و معتبر نتایج پژوهش با کارهای پیشین فراهم شود، و دوم، بررسی عملکرد مدل روی داده‌های متتنوع و واقعی از حوزه‌های مختلف ممکن گردد. این گراف‌ها از منابع مختلفی استخراج شده‌اند و نمایانگر شبکه‌هایی با کاربردهای متتنوع، از جمله شبکه‌های اجتماعی، زیستی، فیزیکی، و اطلاعاتی هستند. همچنین، گراف‌ها از نظر اندازه و ساختار به گونه‌ای انتخاب شدند که مجموعه‌ای جامع و متتنوع ارائه دهنده بطوری که مدل بتواند عملکرد خود را در شرایط مختلف نشان دهد. در جدول ۲-۳ می‌توان اطلاعات گراف‌های واقعی انتخاب شده را دید.

گراف‌های مدل باراباسی-آلبرت نیز همانطور که در بخش تعاریف ذکر شد، به گونه‌ای طراحی شده‌اند که ساختار مشابهی با شبکه‌های دنیای واقعی داشته باشند. استفاده از این گراف‌ها به ما اجازه می‌دهد تا مدل مورد نظر را روی شبکه‌هایی با ویژگی‌های نزدیک به دنیای واقعی آموزش دهیم و نتایج حاصل را به محیط‌های واقعی تعمیم دهیم. در این پژوهش، گراف‌های باراباسی-آلبرت با اندازه‌های مختلف ( $n$ ) و میانگین درجه‌های متفاوت ( $m$ ) تولید شدند تا تنوع کافی برای ارزیابی مدل فراهم شود. در جدول ۱-۳ می‌توان اطلاعات این گراف‌های مصنوعی را دید.

### ۲-۱-۳-۳ ساخت گراف‌های BA

برای تولید این گراف‌ها، از توابع کتابخانه NetworkX استفاده شد. جزئیات و مشخصات آماری هر سه مجموعه از گراف‌های حاضر در جدول ۱-۳ به شرح زیر است:

- در اولین مجموعه، گراف‌هایی با اندازه‌های مختلف  $\{3000, 4000, 5000, 6000, 7000, 8000\}$  استفاده شد. جزئیات و مشخصات آماری هر سه تولید شدند. میانگین درجه گره‌ها ( $m$ ) برای تمامی این گراف‌ها برابر با ۴ در نظر گرفته شد.
- در دومین مجموعه، اندازه گراف ثابت و برابر با  $n = 2000$  بود، اما میانگین درجه گره‌ها مقادیر مختلفی از مجموعه  $\{4, 10, 20\}$  به خود گرفت.
- در سومین مجموعه، اندازه گراف ثابت و برابر با  $n = 4000$  در نظر گرفته شد و میانگین درجه گره‌ها مقادیر  $\{2, 6, 10\}$  را اختیار کرد.

تمام گراف‌های تولیدی در قالب فایل‌های لیست یال‌ها(edges) ذخیره شده و در نام فایل‌ها اطلاعات مربوط به اندازه گراف ( $n$ ) و میانگین درجه ( $m$ ) مشخص شده است تا به سادگی قابل شناسایی باشند.

این جدول ۱-۳ اطلاعات مربوط به خصوصیات هر گراف را نمایش می‌دهد. در ادامه توضیح مختصری درباره هر ستون آورده شده است:

جدول ۱-۳: جدول اطلاعات گراف‌های مصنوعی

Name	<i>n</i>	<i>e</i>	$\langle e \rangle$	$e_{\max}$	$e_{\min}$	<i>c</i>	<i>s</i>	$Gn\%$	$Ge\%$	<i>d</i>
ba_1000_4	1000	1996	3.99	63	2	0.02	1	100.00	100.00	0.00
ba_exp1_3000_4	3000	5996	4.00	142	2	0.01	1	100.00	100.00	0.00
ba_exp1_4000_4	4000	7996	4.00	148	2	0.01	1	100.00	100.00	0.00
ba_exp1_5000_4	5000	9996	4.00	203	2	0.01	1	100.00	100.00	0.00
ba_exp1_6000_4	6000	11996	4.00	313	2	0.01	1	100.00	100.00	0.00
ba_exp1_7000_4	7000	13996	4.00	202	2	0.01	1	100.00	100.00	0.00
ba_exp1_8000_4	8000	15996	4.00	191	2	0.00	1	100.00	100.00	0.00
ba_exp2_2000_10	2000	9975	9.97	180	5	0.02	1	100.00	100.00	0.00
ba_exp2_2000_20	2000	19900	19.90	227	10	0.04	1	100.00	100.00	0.01
ba_exp2_2000_4	2000	3996	4.00	116	2	0.01	1	100.00	100.00	0.00
ba_exp3_4000_10	4000	19975	9.99	250	5	0.01	1	100.00	100.00	0.00
ba_exp3_4000_2	4000	3999	2.00	97	1	0.00	1	100.00	100.00	0.00
ba_exp3_4000_6	4000	11991	6.00	249	3	0.01	1	100.00	100.00	0.00

جدول ۲-۳: جدول اطلاعات گراف‌های واقعی

Name	<i>n</i>	<i>e</i>	$\langle e \rangle$	$e_{\max}$	$e_{\min}$	<i>c</i>	<i>s</i>	$Gn\%$	$Ge\%$	<i>d</i>
arenas-pgp	10680	24316	4.55	205	1	0.27	1	100.00	100.00	0.00
CA-GrQc	5242	14496	5.53	81	1	0.53	355	79.32	92.63	0.00
CA-HepTh	9877	25998	5.26	65	1	0.47	429	87.46	95.50	0.00
ChicagoRegional	12979	20627	3.18	7	1	0.04	1	100.00	100.00	0.00
email	1133	5451	9.62	71	1	0.22	1	100.00	100.00	0.01
faa	1226	2410	3.93	34	1	0.07	1	100.00	100.00	0.00
facebook_combined	4039	88234	43.69	1045	1	0.61	1	100.00	100.00	0.01
figeys	2239	6432	5.75	314	1	0.04	9	99.02	99.78	0.00
ia-crime-moreno	829	1475	3.56	25	1	0.01	1	100.00	100.00	0.00
jazz	198	2742	27.70	100	1	0.62	1	100.00	100.00	0.14
LastFM	7624	27806	7.29	216	1	0.22	1	100.00	100.00	0.00
NS	1461	2742	3.75	34	1	0.69	268	25.94	33.33	0.00
p2p-Gnutella04	10876	39994	7.35	103	1	0.01	1	100.00	100.00	0.00
Peh_edge	2426	16631	13.71	273	1	0.54	148	82.44	96.80	0.01
politician_edges	5908	41729	14.13	323	1	0.39	1	100.00	100.00	0.00
powergrid	4941	6594	2.67	19	1	0.08	1	100.00	100.00	0.00
Stelzl	1706	3191	3.74	95	1	0.01	43	94.67	98.28	0.00
tvshow_edges	3892	17262	8.87	126	1	0.37	1	100.00	100.00	0.00
vidal	3133	6726	4.29	129	1	0.06	210	88.83	95.72	0.00
web-EPA	4271	8909	4.17	175	1	0.07	8	99.58	99.87	0.00

•  $G$ : نام یا شناسه گراف

•  $n$ : تعداد گره‌ها در گراف

• تعداد یال‌ها در گراف  $e$

• میانگین درجه گره‌ها  $\langle e \rangle$

$$\langle e \rangle = \frac{2 \cdot e}{n}$$

• بیشترین درجه گره  $e_{\max}$

• کمترین درجه گره  $e_{\min}$

• میانگین ضریب خوشبندی گره‌ها  $c^{38}$

$$c = \frac{\text{تعداد یال‌های واقعی بین همسایگان}}{\text{تعداد یال‌های ممکن بین همسایگان}}$$

• تعداد زیرگراف‌های جدا از هم  $s$

• درصد گره‌های بزرگ‌ترین مؤلفه  $Gn\%^{39}$

$$Gn\% = \frac{\text{تعداد گره‌های بزرگ‌ترین مؤلفه}}{n} \times 100$$

• درصد یال‌های بزرگ‌ترین مؤلفه  $Ge\%^{40}$

$$Ge\% = \frac{\text{تعداد یال‌های بزرگ‌ترین مؤلفه}}{e} \times 100$$

• چگالی گراف  $d^{41}$

$$d = \frac{2 \cdot e}{n \cdot (n - 1)}$$

<sup>38</sup>Average Clustering Coefficient

<sup>39</sup>Portion of Nodes in the Largest Component

<sup>40</sup>Portion of Edges in the Largest Component

<sup>41</sup>Graph Density

### ۳-۱-۳-۳ اصلاح و یکپارچه‌سازی داده‌های گرافی برای تحلیل

پس از انتخاب و جمع‌آوری گراف‌های مربوطه از منابع معتبر علمی، فرآیند پیش‌پردازش داده‌ها بهمنظور تطابق با نیازهای مدل و آماده‌سازی آن‌ها برای تحلیل‌های بعدی آغاز شد. یکی از چالش‌های اصلی در این مرحله، تنوع فرمت‌های فایل‌های گراف‌ها بود. برخی از فایل‌ها در فرمت edges. بودند، در حالی که برخی دیگر نیازمند تبدیل به فرمت‌های استاندارد و متناسب با فرآیند تحلیل بودند.

برای رفع این مشکل، یک اسکریپت پایتون<sup>۴۲</sup> توسعه داده شد که به‌طور خودکار فایل‌های دارای فرمت متفرقه را خوانده و آن‌ها را به فرمت استاندارد تبدیل می‌کرد. این اسکریپت علاوه بر فرمت‌های غیراستاندارد، برای اصلاح ساختار فایل‌های خراب<sup>۴۳</sup> نیز به‌کار گرفته شد. به عنوان مثال، برخی از فایل‌ها نیاز به تبدیل به فرمت‌های خاصی داشتند که به‌طور دستی در اسکریپت پیاده‌سازی شد تا داده‌ها به درستی با مدل تطبیق پیدا کنند. این فرآیند شامل استخراج داده‌ها از فایل‌های ورودی، اعمال اصلاحات لازم بر ساختار گراف‌ها، و ذخیره‌ی فایل‌ها در فرمت‌های مطلوب بود. همچنین، در خلال این مرحله، داده‌های نامعتبر، یال‌های ناقص و گره‌های غیرمرتب شناسایی و حذف شدند تا از صحت و یکپارچگی داده‌ها اطمینان حاصل گردد. این اقدامات به‌منظور تضمین هماهنگی کامل داده‌ها با نیازهای مدل، هماهنگی با تحقیقات قبلی بر روی این دادگان‌ها و پیشبرد مراحل بعدی پروژه انجام شد.

### ۴-۱-۳-۳ تولید برچسب قدرت انتشار

همان‌طور که در شبه‌کد ۱ نشان داده شده است، مراحل کلی الگوریتم به شکل زیر است:

۱. دریافت گراف از ورودی

۲. محاسبه مقادیر  $\beta$  برای گراف

۳. اجرای مدل SIR برای هر گره با مقادیر مختلف  $\beta$

۴. ذخیره نتایج مربوط به تأثیرگذاری

### ۵-۱-۳-۳ اهمیت برچسب قدرت انتشار و توزیع مقادیر آن‌ها

در فرآیند آموزش مدل‌های یادگیری ماشین، دقت در تولید برچسب‌ها از اهمیت بالایی برخوردار است. به ویژه در پروژه‌ای که در آن هدف شناسایی و رتبه‌بندی گره‌های تأثیرگذار در گراف‌های پیچیده است،

<sup>42</sup>Python

<sup>43</sup>Corrupted File

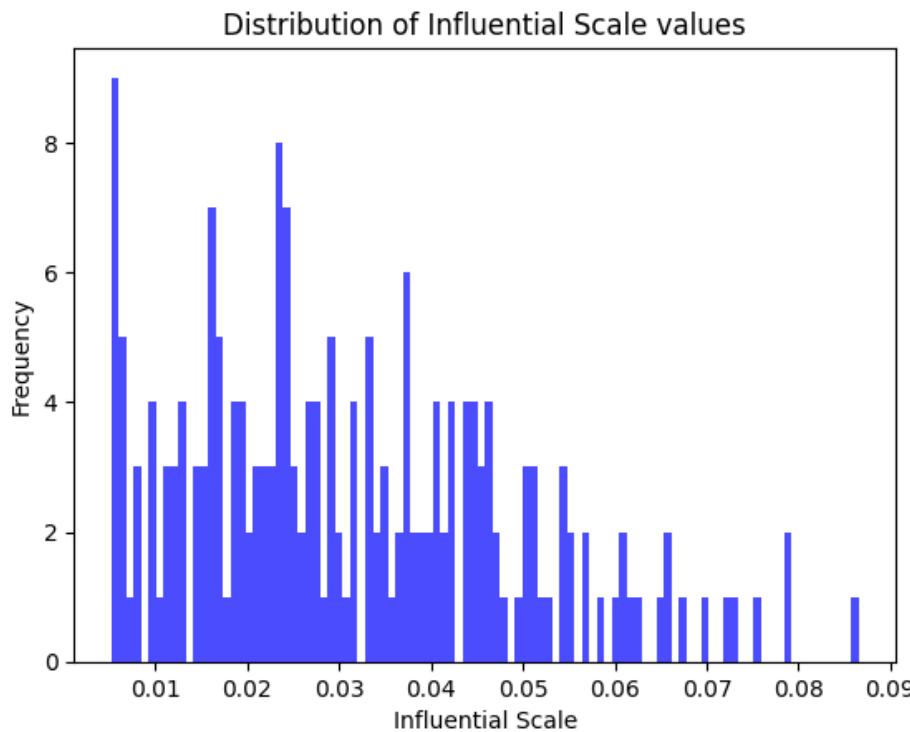
برچسب‌های نادرست می‌توانند تأثیرات منفی و جرماناپذیری بر عملکرد مدل داشته باشند. این امر بهویژه در مواقعی که مقادیر برچسب‌ها بسیار نزدیک به یکدیگر هستند، حائز اهمیت می‌شود. در این پروژه، برچسب‌ها نمایانگر میزان تأثیرگذاری(قدرت انتشار) گره‌ها هستند و حتی کوچک‌ترین تغییر در مقدار آن‌ها می‌تواند به‌طور قابل توجهی ترتیب گره‌ها را تغییر دهد. در حقیقت، اهمیت اصلی در اینجا بر ترتیب برچسب‌ها است، نه بر مقدار دقیق آن‌ها. به عبارت دیگر، ممکن است اختلاف مقدارهای برچسب‌ها تنها یک صد هزارم باشد، اما همین تغییرات اندک می‌تواند منجر به جابه‌جایی قابل توجه گره‌ها در رتبه‌بندی نهایی شود. اگر برچسبی به اشتباه بزرگ‌تر از برچسب گره‌ای دیگر باشد، این موجب تغییر در رتبه‌بندی نهایی می‌شود، چرا که مدل به اشتباه این گره را به عنوان گره‌ای با تأثیر بیشتر شناسایی می‌کند. در این شرایط، حتی یک تغییر کوچک در مقدار برچسب می‌تواند ترتیب نهایی گره‌ها را به‌طور کامل دگرگون کند، که این امر می‌تواند کیفیت پیش‌بینی‌ها و نتایج مدل را به‌طور چشمگیری کاهش دهد.

بنابراین، دقت در تولید برچسب‌ها و اطمینان از صحت آن‌ها برای حفظ صحت و اعتبار مدل بسیار حیاتی است. در این پروژه، با توجه به حساسیت زیاد مدل به تغییرات کوچک در برچسب‌ها، تمرکز بر تولید برچسب‌های دقیق و صحیح امری ضروری است. یک نمونه از نمودار ستونی<sup>۴۴</sup> توزیع برچسب‌های قدرت انتشار که در این پروژه استخراج شده‌اند، در شکل ۲-۳ آمده است. در این نمودار، محور افقی مقادیر برچسب‌ها را نشان می‌دهد و محور عمودی نیز نشان‌دهنده فراوانی هر یک از مقادیر در داده‌ها است. همانطور که در تصویر مشاهده می‌شود، برچسب‌ها دارای توزیع خاصی هستند که در تحلیل‌های بعدی به آن پرداخته خواهد شد. این نمودار ستونی به‌وضوح نشان می‌دهد که مقادیر برچسب‌ها در نزدیکی یکدیگر قرار دارند و تغییرات کوچک در مقادیر آن‌ها می‌تواند تأثیر زیادی بر ترتیب گره‌ها در مدل نهایی بگذارد. بهویژه با توجه به نزدیکی مقادیر، حتی اختلافهای اندک می‌توانند موجب تغییر در نتایج رتبه‌بندی شوند که در نهایت تأثیر زیادی بر پیش‌بینی‌های مدل دارد.

### ۳-۱-۶ تأثیر پارامتر $\beta$ روی برچسب‌های تولیدی

. پارامتر  $\beta$ ، که به عنوان احتمال انتقال بیماری (یا اطلاعات) بین گره‌های شبکه تعریف می‌شود، تأثیر مستقیمی بر رفتار انتشار دارد و به شدت تحت تأثیر ساختار شبکه و توزیع درجه گره‌ها قرار می‌گیرد [۷، ۸]. در این بخش، تأثیر مقادیر مختلف  $\beta$  بر نتایج نهایی و برچسب‌های تولیدی تحلیل می‌شود. تحقیقات پیشین [۷، ۸] حاکی از آن است که در حوالی  $\beta_c$ ، تغییرات اندک در مقدار  $\beta$  می‌تواند

<sup>44</sup>Histogram



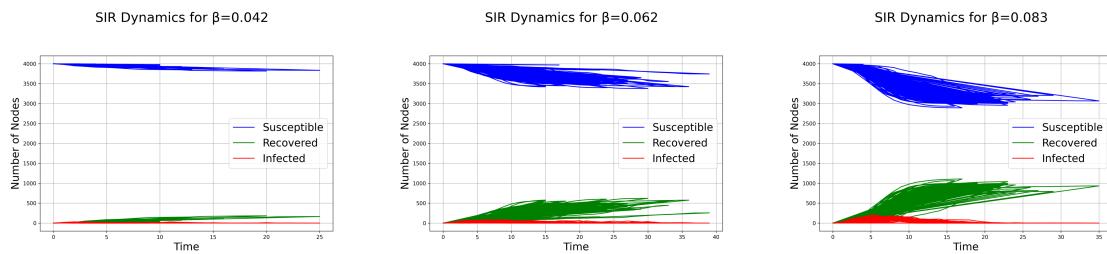
شکل ۳-۲: توزیع مقادیر برچسبها در گراف jazz

تأثیرات چشمگیری بر تعداد گرهای آلوده داشته باشد. این در واقع همان آستانه‌ی انتشار است که در قسمت تعاریف ذکر شد.

این تحقیقات، همچنین نشان داد که در مقادیر پایین  $\beta$ ، بخش بزرگی از گرهها آلوده نمی‌شوند و انتشار محدود به یک ناحیه کوچک باقی می‌ماند. در حالی که در مقادیر بالا، کل شبکه به سرعت آلوده می‌شود و قدرت تمایز میان تأثیر گرهها کاهش می‌یابد. این موضوع در شکل ۳-۳ به تصویر کشیده شده است. این یافته‌ها به بهینه‌سازی انتخاب مقادیر  $\beta$  برای تحلیل‌های مختلف کمک کرده و نشان می‌دهند که برای شناسایی گرهای تأثیرگذار، انتخاب مقادیر نزدیک به  $c\beta$ ، بهترین نتایج را ارائه می‌دهند، در حالی که برای بررسی انتشار کلی، مقادیر بالاتر  $\beta$  مفیدتر هستند. بنابراین، انتخاب مقادیر مناسب  $\beta$  برای تحلیل تأثیرات در شبکه‌های پیچیده، نیازمند دقت بالایی است و می‌تواند به طور مستقیم بر کیفیت و دقت برچسب‌های تولیدی تأثیر بگذارد.

در این پژوهش، با استفاده از نظریه میدان میانگین و معادله ۳-۲-۲-۲، مقدار آستانه‌ای  $c\beta$  برای هر شبکه محاسبه شده و مقادیر  $\beta$  به صورت مضری از این مقدار آستانه‌ای در نظر گرفته شدند.

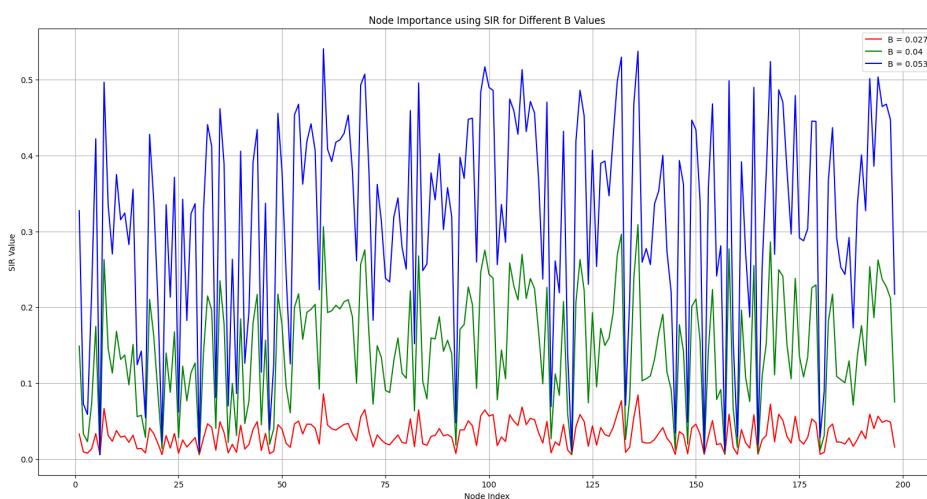
شکل ۳-۳ روند تغییرات تعداد گرهای مستعد، آلوده و بهبود یافته را در طول زمان برای مقادیر مختلف  $\beta$  نشان می‌دهد. در ابتدای شبیه‌سازی، تعداد گرهای مستعد زیاد است و به تدریج با افزایش



شکل ۳-۳: روند تغییرات تعداد گره‌های مستعد ( $S$ )، آلوده ( $I$ ) و بهبود یافته ( $R$ ) در طول زمان برای مقادیر مختلف  $\beta$

**تعداد گره‌های آلوده کاهش می‌یابد.** این کاهش با نرخ  $\beta$  رابطه مستقیمی دارد؛ به طوری که هرچه  $\beta$  بزرگ‌تر باشد، کاهش گره‌های مستعد سریع‌تر خواهد بود. تعداد گره‌های آلوده معمولاً به شکل یک قله است، به این معنی که در آغاز شبیه‌سازی رشد می‌کند، به اوج می‌رسد و سپس کاهش می‌یابد. زمان رسیدن به این اوج و مقدار آن به مقدار  $\beta$  وابسته است، به طوری که  $\beta$  بیشتر منجر به اوج سریع‌تر و شدیدتر می‌شود. در همین حال، تعداد گره‌های بهبود یافته در طول زمان به صورت پیوسته افزایش یافته و در نهایت به یک مقدار ثابت می‌رسد، که نشان‌دهنده دامنه نهایی انتشار<sup>۴۵</sup> است. این تحلیل به وضوح تأثیر پارامتر  $\beta$  بر سرعت و شدت انتشار بیماری را نشان می‌دهد و به ارزیابی رفتار شبکه در شرایط مختلف کمک می‌کند.

برای درک تأثیر  $\beta$  بر روی مقادیر برچسب‌های هر گره (قدرت انتشار) نیز می‌توان به شکل ۴-۳ اشاره کرد. از آنجایی که معادله تولید برچسب<sup>۱-۳</sup> رابطه‌ی مستقیم با تعداد گره‌های آلوده و بهبود یافته دارد، افزایش مقدار  $\beta$  منجر به افزایش مقدار برچسب تولیدی برای گره‌ها می‌شود.



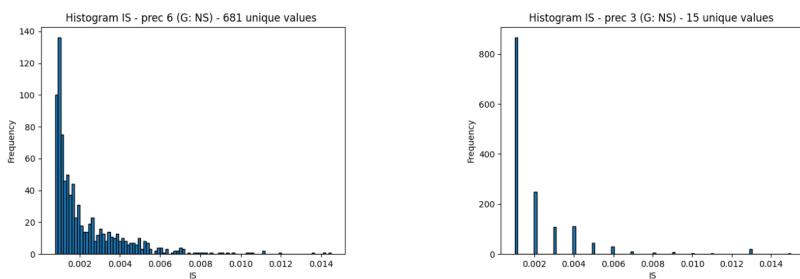
شکل ۴-۳: تحلیل اهمیت گره‌ها در شبیه‌سازی SIR برای مقادیر مختلف  $\beta$

<sup>45</sup>Final Spread Range

### ۳-۱-۷ تأثیر دقت‌های مختلف در شبیه‌سازی

در این بخش، تأثیر دقت‌های مختلف در برچسب‌های تولید شده بررسی می‌شود. استفاده از دقت‌های پایین‌تر، به‌ویژه زمانی که مقادیر برچسب‌ها کوچک هستند و تغییرات آن‌ها در ارقام اعشاری بسیار دقیق مانند ۵ رقم بعد از اعشار رخ می‌دهد، ممکن است تأثیرات قابل توجهی در نتایج داشته باشد. هنگامی که برچسب‌ها را تنها با سه رقم بعد از اعشار ذخیره می‌کنیم، ممکن است مقدار زیادی از گره‌ها اهمیت یکسانی پیدا کنند، زیرا دقت کافی برای تمایز دقیق بین مقادیر برچسب‌ها وجود نخواهد داشت. این موضوع به‌ویژه در پروژه‌هایی که ترتیب اهمیت گره‌ها مهم است، می‌تواند مشکل‌ساز شود. به عنوان مثال، در این پروژه، که رتبه‌بندی گره‌ها بر اساس قدرت انتشار آن‌ها در شبکه اهمیت زیادی دارد، بسیاری از گره‌ها ممکن است به دلیل فقدان دقت مناسب در مقادیر برچسب‌ها، در یک موقعیت مشابه در ترتیب اهمیت قرار گیرند، و این باعث کاهش کیفیت پیش‌بینی‌ها و کاهش دقت مدل خواهد شد. در این حالت، انتخاب دقت مناسب برای ذخیره‌سازی مقادیر برچسب‌ها می‌تواند به‌طور چشمگیری بر کیفیت نتایج نهایی تأثیر بگذارد.

در این بخش، به تحلیل دقیق‌تر تأثیر دقت‌های مختلف بر نتایج شبیه‌سازی پرداخته می‌شود. برای داده‌هایی که با دقت ۳ رقم بعد از اعشار ذخیره شده‌اند، نمودار ستونی **۵-۲** نشان می‌دهد که اغلب مقادیر در یک بازه کوچک تجمع یافته‌اند. به‌ویژه، اولین میله نمودار که مربوط به فرکانس مقادیر مشابه است، فرکانسی بیش از  $80^{\circ}$  را نمایش می‌دهد. این موضوع به وضوح نشان‌دهنده این است که اختلافات جزئی بین قدرت انتشار، که تنها در ارقام ۴ تا ۶ بعد از اعشار قابل شناسایی هستند، به‌طور کامل نادیده گرفته شده و تمام گره‌ها با مقادیر مشابه در یک ستون قرار گرفته‌اند. این تجمع مقادیر به یکدیگر به‌طور خاص در مسائل رتبه‌بندی گره‌ها مشکل‌ساز است، چرا که در چنین شرایطی بسیاری از گره‌ها با مقادیر مشابه در نظر گرفته می‌شوند و اهمیت آن‌ها به‌طور نادرست مشابه در نظر گرفته می‌شود. در مقابل، هنگامی که دقت به ۶ رقم بعد از اعشار افزایش می‌یابد، توزیع مقادیر در نمودار ستونی به‌طور قابل توجهی تغییر می‌کند. در این حالت، هیچ ستونی در نمودار از فرکانس  $140^{\circ}$  فراتر نمی‌رود و مقادیر به‌طور یکنواخت‌تر و پخش‌تری در بازه‌های مختلف توزیع می‌شوند. این تغییر نشان‌دهنده حفظ تفاوت‌های جزئی بین گره‌ها است که به‌ویژه در این پروژه و دیگر پروژه‌های حساس به ترتیب، نقشی حیاتی ایفا می‌کند. دقت بالاتر باعث می‌شود که تمایزات ظریف میان گره‌ها حفظ شود و این موضوع موجب بهبود دقت و صحت نتایج مدل می‌گردد. همانطور که در نمودار ستونی مشاهده می‌شود، مقادیر قدرت انتشار برای دقت ۳ رقم بعد از اعشار (سمت راست) و دقت ۶ رقم بعد از اعشار (سمت چپ) تفاوت زیادی دارند.



شکل ۳-۵: اثر دقت مقادیر انتشار قدرت و توزیع آن در دقت‌های مختلف

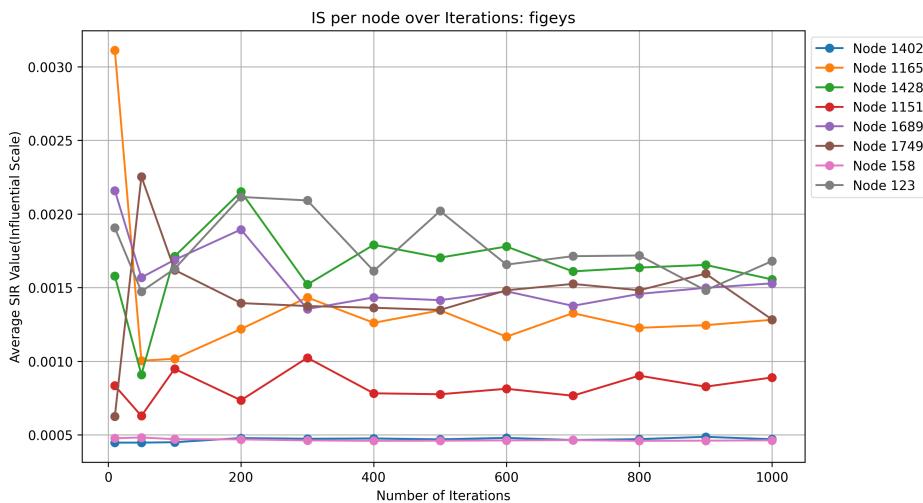
### ۳-۱-۸ تأثیر تعداد تکرار شبیه‌سازی

در این بخش از گزارش، به تأثیر تعداد تکرار شبیه‌سازی‌های SIR بر روی برچسب‌های تولیدی می‌پردازیم. برای شبیه‌سازی انتشار بیماری (یا اطلاعات) در شبکه، هر گره‌ای که مورد بررسی است، به عنوان «آلوده» یا «بیمار» فرض می‌شود و فرآیند شبیه‌سازی از این گره آغاز می‌شود. برای هر بار شبیه‌سازی، وضعیت گره‌ها در شبکه بر اساس مدل SIR به روزرسانی می‌شود. در این فرآیند، تعداد دفعات تکرار شبیه‌سازی نقش مهمی در دقت نتایج ایفا می‌کند. هرچه تعداد دفعات شبیه‌سازی بیشتر باشد، میانگین مقادیر بدست آمده از انتشار بیماری برای گره‌های مختلف دقیق‌تر خواهد بود. این به این معناست که با افزایش تعداد تکرارها، داده‌ی پرت<sup>۴۶</sup> کمتری در مقادیر قدرت انتشار مشاهده می‌شود و نتایج به واقعیت نزدیک‌تر خواهد بود. در واقع، هرچه تعداد تکرارها بیشتر باشد، دقت در محاسبه مقادیر قدرت انتشار افزایش می‌یابد، و تغییرات تصادفی در فرآیند شبیه‌سازی کاهش پیدا می‌کند.

این نکته که افزایش تعداد تکرارها باعث افزایش زمان محاسبات می‌شود نیز حائز اهمیت است. بنابراین، باید تعادلی میان دقت نتایج و زمان محاسبات برقرار کرد. در پروژه‌ای که در دست داریم، که ترتیب گره‌ها و روابط آن‌ها در شبکه اهمیت زیادی دارد، دقت در شبیه‌سازی‌ها برای استخراج قدرت انتشار صحیح بسیار مهم است. به عبارت دیگر، در شرایطی که تعداد تکرارها افزایش یابد، اگرچه زمان بیشتری صرف می‌شود، اما نتیجه نهایی با نویز کمتری ارائه خواهد شد و می‌توان به تحلیل‌های دقیق‌تری دست یافت. در این پروژه، تلاش می‌شود تا تعداد تکرارها به گونه‌ای تنظیم شود که از یک سو کمترین نویز را داشته باشیم و از سوی دیگر زمان پردازش نیز در حد معقولی باقی بماند.

نمودار ۳-۶ نشان می‌دهد که چگونه میانگین مقادیر قدرت انتشار برای ۸ گره تصادفی در شبکه با افزایش تعداد تکرارهای شبیه‌سازی تغییر می‌کند. به طور خاص، با افزایش تعداد تکرارها، دقت محاسبه مقادیر قدرت انتشار بهبود می‌یابد و نویز کمتری در نتایج مشاهده می‌شود. در این نمودار، محور افقی

<sup>46</sup>Noise



شکل ۳-۶: تأثیر تعداد تکرار شبیه‌سازی بر روی قدرت انتشار ۸ گرهای تصادفی در گراف Figeys

تعداد تکرارهای شبیه‌سازی را نشان می‌دهد و محور عمودی میانگین مقادیر قدرت انتشار برای هر گره را نمایش می‌دهد. همان‌طور که مشاهده می‌شود، با افزایش تعداد تکرارها، مقدار قدرت انتشار به تدریج پایدارتر و دقیق‌تر می‌شود، که نشان‌دهنده تاثیر مثبت افزایش تعداد تکرارها بر دقت نتایج است. در این شبیه‌سازی‌ها، ۸ گره تصادفی از شبکه انتخاب و برای هر کدام فرآیند شبیه‌سازی با تعداد متفاوتی از تکرارها (تعداد تکرارهای [۱۰, ۵۰, ۱۰۰, ۲۰۰, ۳۰۰, ۴۰۰, ۵۰۰, ۱۰۰۰] بر روی خط افقی قبل مشاهده است) اجرا شده است. نتایج به دست آمده از این شبیه‌سازی‌ها به وضوح نشان می‌دهند که تعداد تکرارهای بیشتر به بهبود دقت نتایج کمک می‌کند و قدرت انتشار گرهات به شکلی معنادار و دقیق‌تر محاسبه می‌شود.

### ۹-۱-۳-۳ تحلیل زمان اجرای ساخت برچسب‌ها

برای ایجاد برچسب‌های مرتبط با شبیه‌سازی، نیاز به انجام تعداد مشخصی از تکرارها برای هر گره داریم تا نوسانات ناشی از شبیه‌سازی‌های تصادفی کاهش یابد و داده‌های پرت کمتری به دست آید. در این پژوهش، برای هر گره در گراف، ۱۰۰۰ تکرار از شبیه‌سازی SIR انجام شده و میانگین‌گیری بر نتایج حاصل صورت گرفته است. این فرآیند تضمین می‌کند که نتایج نهایی از پایداری و اعتبار لازم برخوردار باشند. این تعداد تکرار موجب افزایش دقت نتایج شده اما به شدت زمان محاسبات را افزایش داده است. برای نمونه، میانگین زمان موردنیاز برای تولید برچسب برای یک گره (و کل گراف) به ازای تعداد تکرارهای مختلف در گرافی از مدل مدل برابسی-آلبرت، با ۷۰۰۰ گره و میانگین ۴ یال در جدول ۴-۳ بررسی شده است.

همان‌طور که در جدول ۴-۳ مشاهده می‌شود، برای تولید برچسب‌های مرتبط با تنها یک گره از گراف

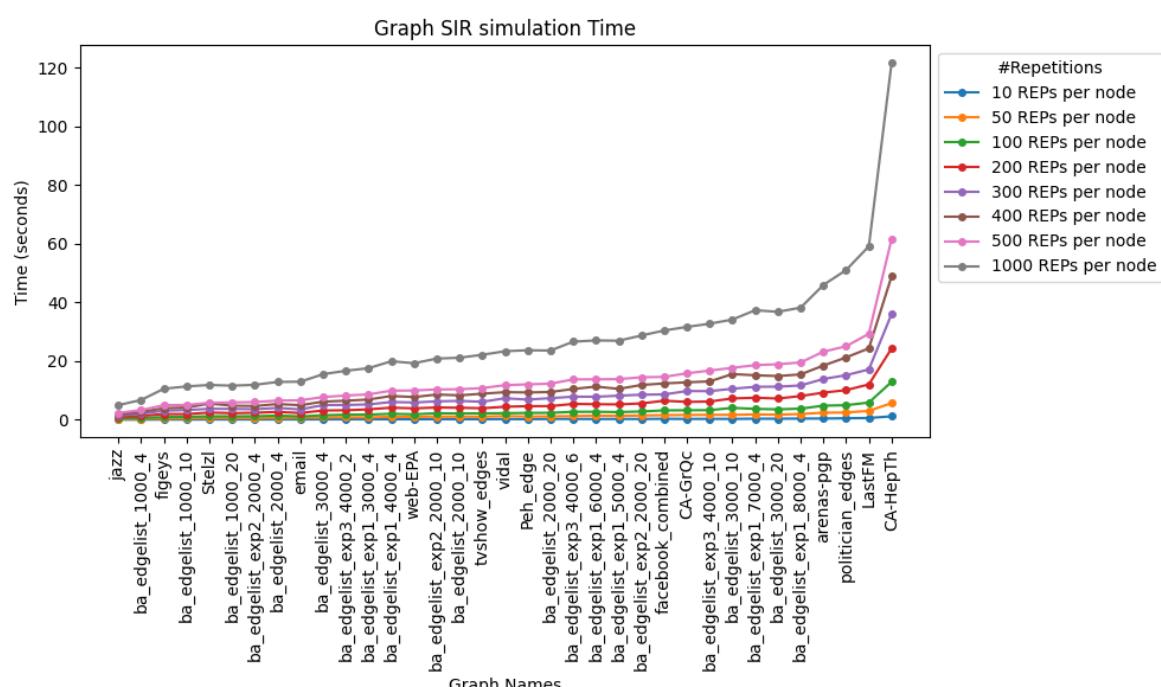
## جدول ۳-۳: نمونه خروجی استخراج ویژگی‌های حجمی

جدول ۴-۳: زمان محاسبه برای تولید برچسب برای یک گره در گراف ۴-۰۰۰ به ازای ba\_edgelist\_exp1\_7000\_4 تعداد تکرارهای مختلف.

۱۰	۵۰	۱۰۰	۲۰۰	۳۰۰	۴۰۰	۵۰۰	۱۰۰۰	برای یک گره	برای کل گراف
۰۳۷(s) ۱(hr)	۱۸۳(s) ۳(hr)	۲۷۲(s) ۷(hr)	۷۵۲(s) ۱۴(hr)	۱۱۲۴(s) ۲۱(hr)	۱۵۱۷(s) ۲۹(hr)	۱۸۶۲(s) ۳۶(hr)	۳۷۳۴(s) ۷۳(hr)		

مذکور با ۱۰۰۰ تکرار شبیه‌سازی، به طور میانگین حدود ۷۲ ساعت (۳ روز) زمان لازم است. این زمان تنها برای ساخت برچسب‌های یک گراف و تحت فرض استفاده از یک مقدار ثابت برای نرخ انتقال محاسبه شده است. با این حال، در عمل برای شبیه‌سازی‌های دقیق‌تر، باید از مقادیر مختلفی از نرخ انتقال استفاده شود تا رفتار گراف در شرایط مختلف بررسی گردد. بنابراین، زمان کلی محاسبه برای این گراف باید در تعداد مقادیر  $\beta$  ضرب شود تا مدت زمان واقعی مورد نیاز به دست آید.

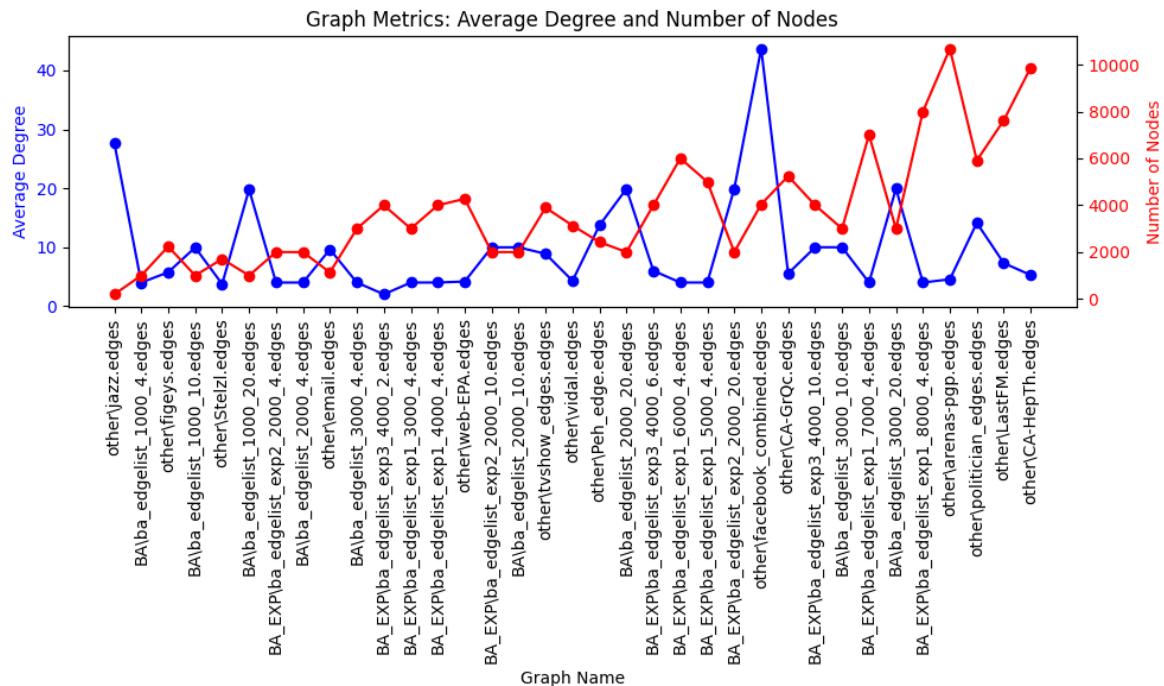
علاوه بر این، باید توجه داشت که این گراف حتی زمان برترین گراف در بین گراف‌های ما نیست. همانطور که در شکل ۷-۳ مشاهده می‌کنید، گراف‌های با اندازه‌های بزرگ‌تر و ساختار پیچیده‌تر، زمان محاسبات را به میزان قابل توجهی افزایش می‌دهند. این موضوع نشان‌دهنده اهمیت بهینه‌سازی زمان محاسبات با استفاده از روش‌هایی مانند موازی‌سازی یا افزایش تعداد محیط‌های شبیه‌سازی است.



شکل ۷-۳: زمان شبیه‌سازی گراف‌های مختلف

شکل ۸-۳ همین گراف‌ها را با ترتیب قبلی (افزایش زمان شبیه‌سازی) به همراه میانگین درجه‌ی

گره‌ها و تعداد گره‌ها به نمایش کشیده است. اگرچه زمان شبیه‌سازی وابسته به ساختار و ویژگی‌های ساختاری گراف است، اما همانطور که مشاهده می‌شود، ویژگی‌های ذکر شده نیز تاثیر زیادی روی این زمان دارند.

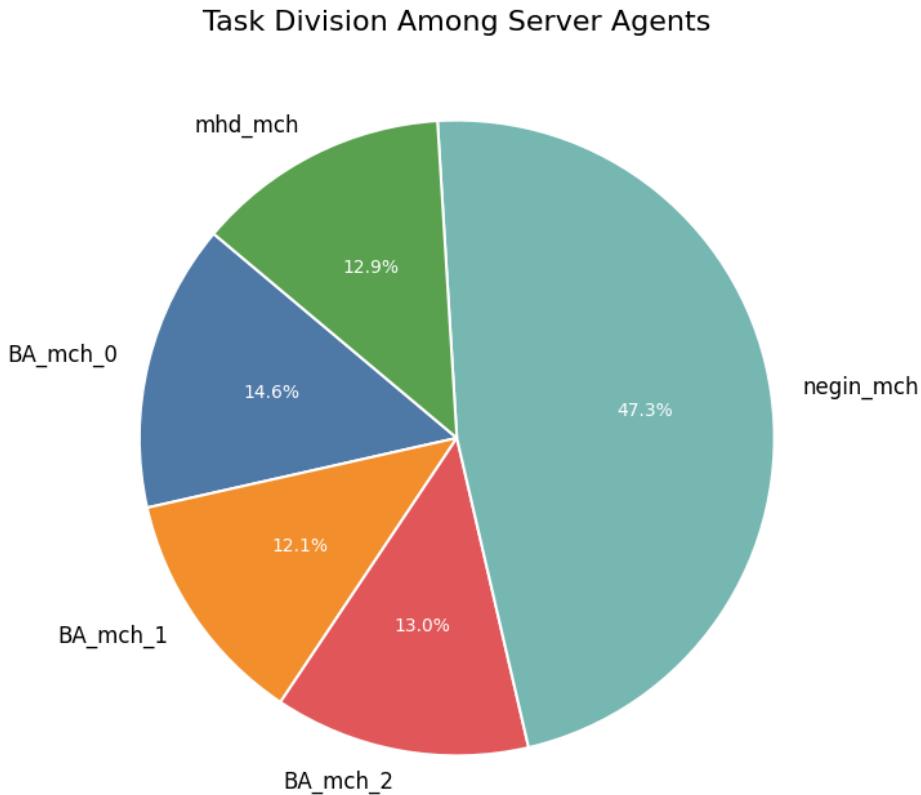


شکل ۳-۸: تأثیر ترکیبی میانگین درجه‌ی گرفه‌ها و تعداد گرفه‌ها بر افزایش زمان شبیه‌سازی

### ۱۰-۱-۳-۳ تسریع فرآیند شبیه‌سازی

همانطور که در قسمت قبل به آن اشاره شد، یکی از چالش‌های اصلی در فرآیند شبیه‌سازی، زمان طولانی محاسبات برای گراف‌های بزرگ و تعداد تکرارهای بالا است. همانطور که در بخش قبلی اشاره شد، این زمان با افزایش تعداد مقادیر  $\beta$  یا گراف‌های بزرگ‌تر به‌طور قابل توجهی افزایش می‌یابد. برای حل این مشکل، فرآیند شبیه‌سازی بهینه‌سازی و به صورت کاملاً موازی بازنویسی شد.

سپس برای تقسیم کار راحت تر، کد شبیه‌سازی داکریزه شده و در قالب کانتینرهای مستقل بهینه‌سازی شد. این کار امکان اجرای کد روی سرورهای مختلف با تنظیمات یکسان را فراهم کرد. در مرحله بعد، شبیه‌سازی‌ها بر روی ۵ دستگاه سرور مجزا توزیع شدند. هر دستگاه به گونه‌ای پیکربندی شد که بتواند چندین شبیه‌سازی را به صورت همزمان با استفاده از multithreading انجام دهد. این امکان‌پذیر بود زیرا شبیه‌سازی‌های مختلف به یکدیگر وابستگی نداشتند و کاملاً مستقل از یکدیگر اجرا می‌شدند.



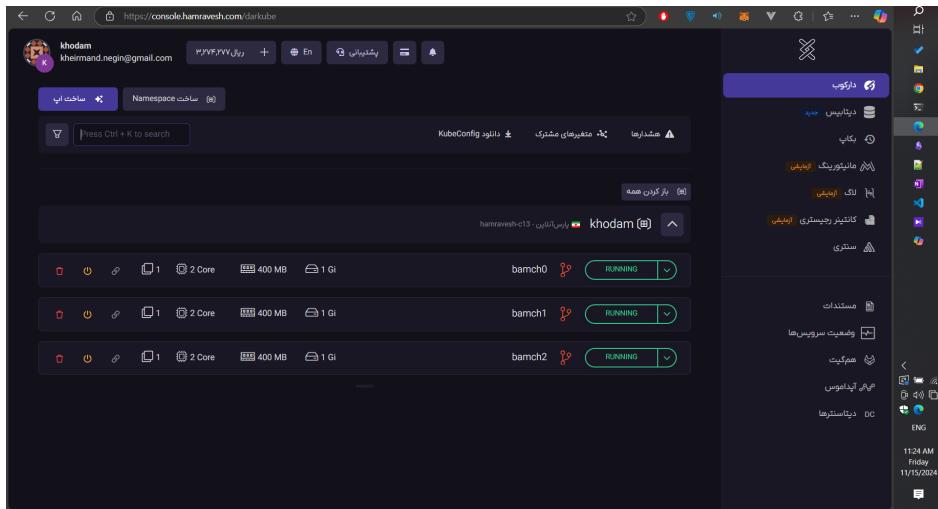
شکل ۹-۳: ساختار تقسیم وظایف میان سرورها

فرآیند توزیع وظایف میان سرورها به صورت این صورت طراحی شد:

- گره‌های هر گراف به تعداد سرورهای در دسترس و میزان توان محاسبیتی آن‌ها تقسیم شدند.
- هر سرور بخشی از گره‌ها را برای شبیه‌سازی دریافت کرد.
- در هر سرور، فرآیندهای شبیه‌سازی به کمک چندین thread به‌طور موازی انجام می‌شد.

این طراحی به‌گونه‌ای پیاده‌سازی شد که از تمام هسته‌های پردازنده هر سرور استفاده بهینه شود. شکل ۹-۳ ساختار تقسیم وظایف میان سرورها و نحوه تخصیص گره‌ها به فرآیندهای موازی را نشان می‌دهد. سرورهای BA\_mch\_1، BA\_mch\_0، BA\_mch\_2 همه، ماشین‌های مجازی هستند که با استفاده از «همروش»<sup>۴۷</sup> (یک بستر ابری استقرار نرم‌افزار) برای پروژه در اختیار گرفته و شروع به فعالیت کردند. شکل ۱۰-۳ نمایی از این بستر را نشان میدهد.

<sup>47</sup> <https://hamravesh.com/>



شکل ۳-۱۰: نمایی از هموروش

این تغییرات باعث شدند که زمان کل محاسبات به طور چشم‌گیری کاهش یابد. برای مثال، شبیه‌سازی یک گراف بزرگ که قبلاً نیازمند چندین روز زمان بود، اکنون با استفاده از این رویکرد در کمی بیشتر از یک روز تکمیل می‌شد.

### ۱۱-۱-۳-۳ یکپارچه‌سازی و سازماندهی فایل‌های تولید شده توسط شبیه‌سازی

در این پژوهه، به دلیل اینکه برنامه‌ی تولید قدرت‌های انتشار روی چند دستگاه به‌طور همزمان در حال اجرا بود، نیاز به پردازش و ترکیب<sup>۴۸</sup> خروجی‌های تولید شده از تمامی دستگاه‌ها و سپس رفع مشکلات احتمالی در فایل‌ها وجود داشت. این مرحله به منظور تضمین هماهنگی و یکپارچگی داده‌ها<sup>۴۹</sup> و حذف هرگونه خطا یا ناهمخوانی انجام شد. مشکلاتی که در فایل‌های تولید شده پیش آمد و مسئولیت رفع آن ها با برنامه‌ی یکپارچه‌سازی دادگان بود، به شرح زیر است:

۱. ساخت چندین فایل CSV برای هر مقدار<sup>۵۰</sup>: برای هر مقدار از نرخ انتقال، یک فایل CSV مجزا ایجاد می‌شد که برچسب‌های مربوط به آن مقدار در آن ذخیره می‌شد. بنابراین، تعداد سطرهای هر فایل CSV برابر با تعداد گره‌ها است. اگر این شرط برقرار نمی‌شد و تعداد سطرها برای هر فایل مختلف بود، باید گره‌هایی که در فایل‌ها وجود نداشتند، مجدداً شبیه‌سازی می‌شدند.

۲. حذف تکرارها و محاسبه میانگین: در صورتی که برای یک گره در یک فایل CSV، چندین مقدار برای قدرت انتشار وجود داشت، برای حفظ دقت و یکپارچگی داده‌ها، میانگین این مقادیر محاسبه

<sup>48</sup>Merge

<sup>49</sup>Data Consistency

می‌شد و مقدار نهایی به فایل اضافه می‌شد. سپس سطرهای تکراری حذف می‌شدند تا فایل به صورت مرتب و بدون خطای باقی بماند.

۳. مرتب‌سازی فایل‌ها<sup>۵۰</sup>: برای اینکه داده‌ها در فایل‌ها به صورت مرتب و خوانا باشند، فایل‌های CSV بر اساس شناسه گره‌ها مرتب‌سازی می‌شوند. این عملیات به منظور تمیز کردن و ایجاد نظم در داده‌ها انجام می‌شد تا پس از پردازش، کاربر با استفاده از رابط کاربری بتواند به راحتی به نتایج دست یابد.

۴. رفع مشکلات موجود در فایل‌ها: در برخی موارد ممکن بود مشکلاتی در ساختار فایل‌ها پیش آید، مانند عدم هم‌راستایی داده‌ها. این مشکلات با استفاده از ابزارهای پردازش داده‌ها رفع می‌شوند تا فایل‌ها برای تجزیه و تحلیل آماده شوند. این عملیات تمیز کردن داده<sup>۵۱</sup> و یکپارچه‌سازی داده‌ها، به عنوان گامی ضروری در پردازش داده‌ها، باعث شد که نتایج نهایی دقیق‌تر و قابل اعتماد‌تر باشند و بتوان تحلیل‌های بیشتری از داده‌های شبیه‌سازی شده استخراج کرد.

---

<sup>50</sup>Sort

<sup>51</sup>Data Cleaning

## فصل چهارم

### سامانه توسعه یافته

در این فصل، روشی که در این پژوهه برای شناسایی گره‌های تأثیرگذار استفاده شده است، به تفصیل شرح داده می‌شود. این فصل به بررسی شبکه عصبی پیچشی گرافی و مراحل اصلی الگوریتم پرداخته و در نهایت، جزئیات رابط کاربری توسعه یافته برای این سامانه را توضیح می‌دهد.

## ۱-۴ شبکه عصبی پیچشی گرافی

در این بخش، به بررسی مفصل‌تری از شبکه عصبی پیچشی گرافی خواهیم پرداخت. این یک روش پیشرفته در یادگیری ماشین است که به طور خاص برای داده‌های گرافی طراحی شده است. برخلاف شبکه‌های عصبی معمولی که بر روی داده‌های ساختار یافته<sup>۱</sup> (داده‌ی متوالی<sup>۲</sup> یا تصویری<sup>۳</sup>) کار می‌کنند، گراف‌ها ساختار غیر خطی دارند که نیازمند رویکردهای خاصی برای پردازش و استخراج ویژگی‌ها می‌باشد. در این بخش، ابتدا به معرفی مدل مورد استفاده، سپس به شرح الگوریتم و کاربرد آن در پژوهه خواهیم پرداخت.

### ۱-۱-۴ معرفی مدل

در مدل‌های سنتی شبکه عصبی پیچشی، عملیات هم‌گشت<sup>۴</sup> به منظور استخراج ویژگی‌های محلی از داده‌های تصویری یا متوالی انجام می‌شود. در گراف‌ها، داده‌ها به صورت گره‌ها و یال‌ها سازماندهی می‌شوند، و گراف‌ها می‌توانند ساختارهای پیچیده‌ای داشته باشند که نیازمند تحلیل‌های خاص است. برای حل این مشکل، شبکه عصبی پیچشی گرافی طراحی شده است که قادر است ویژگی‌های محلی و ساختاری گراف‌ها را از طریق عملیات هم‌گشت به دست آورد.

در این مدل، به جای اعمال عملیات هم‌گشت در فضای متعارف (مانند تصاویر)، این عملیات به گراف‌ها تعمیم داده می‌شود. شبکه‌های عصبی کانولوشنی گرافی به طور مؤثر ویژگی‌ها را از گره‌ها و یال‌های گراف استخراج کرده و روابط ساختاری بین گره‌ها را مدل‌سازی می‌کنند.

<sup>1</sup> Structured Data

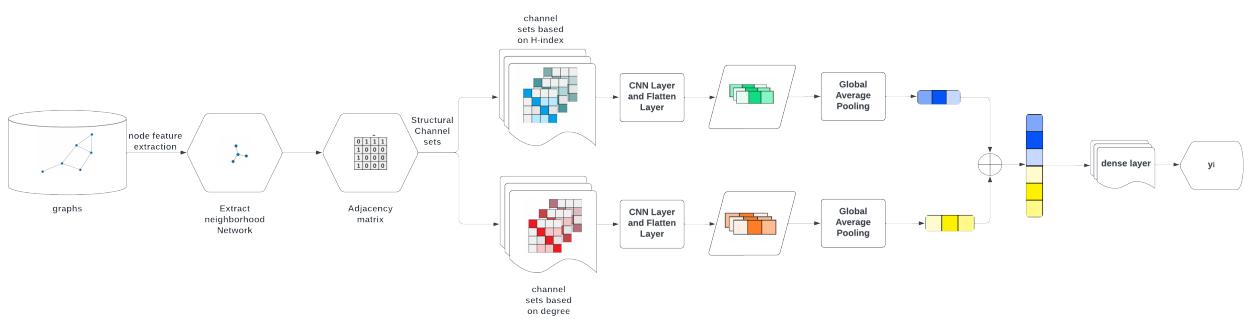
<sup>2</sup> Sequential Data

<sup>3</sup> Image Data

<sup>4</sup> Convolution

## ۲-۱-۴ مراحل اصلی الگوریتم

مراحل اصلی این روش شامل استخراج ویژگی‌های گره، ایجاد مجموعه‌های کanal ساختاری بازنمایی گره، و پیش‌بینی مدل می‌باشد. نمایی از این مراحل در شکل ۱-۴ قابل مشاهده است.



شکل ۱-۴: مدل توسعه یافته

## ۱-۲-۱-۴ استخراج ویژگی‌های گراف

ویژگی‌هایی که استخراج می‌شوند، برخلاف روش‌های سنتی، به‌طور خاص برای تحلیل ساختار شبکه‌های پیچیده و رتبه‌بندی گره‌ها طراحی شده‌اند. این ویژگی‌ها علاوه بر آنکه بر پایه‌ی ویژگی‌های کلاسیک مانند درجه گره و همسایگان آن ساخته شده‌اند، از رویکردهایی استفاده می‌کنند که قادرند ویژگی‌های غیرمستقیم و پیچیده‌تر شبکه را نیز در نظر بگیرند. در واقع، آن‌ها با در نظر گرفتن اثرات تجمعی و تأثیرات غیرمستقیم به شبیه‌سازی دقیق‌تری از وضعیت گره‌ها و شبکه دست می‌یابند. ویژگی‌های موردنظر به دو دسته کلی تقسیم می‌شوند: ویژگی‌های مبتنی بر مرکزیت درجه ( $W^D$ ) و ویژگی‌های مبتنی بر شاخص  $H$  ( $W^H$ ). هر دو دسته برای کاهش پیچیدگی محاسباتی انتخاب شده‌اند. هر دسته شامل سه نوع ماتریس است که بر اساس اطلاعات همسایگی مستقیم (یک گام)، همسایگی تجمعی (دو گام) و همسایگی تعمیم‌یافته (سه گام) تعریف شده‌اند. این کار به‌منظور افزایش دقت انجام شده تا علاوه بر تأثیر مستقیم گره، تعداد ارتباطات و مرکزیت درجه همسایگان نیز در تحلیل گره تأثیر بگذارند.

## ۲-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی مستقیم ( $W^{D1}$ )

این ماتریس تعداد اتصالات مستقیم گره با همسایگان خود را نشان می‌دهد. مقدار این ماتریس برای گره  $i$  به صورت زیر تعریف می‌شود:

$$W_i^{D1} = \sum_{j \in N_i} a_{ij}$$

که در آن:

$N_i$  • مجموعه همسایگان گره  $i$

- $a_{ij}$  • مقدار خانه‌ی  $(i, j)$  ام از ماتریس مجاورت<sup>5</sup> که نشان‌دهنده وجود یا عدم وجود یال بین گره  $i$  و  $j$  است. اگر یال وجود داشته باشد، مقدار آن ۱ و در غیر این صورت ۰ است.

#### ۳-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی تجمعی ( $W^{D2}$ )

این ماتریس علاوه بر اتصالات مستقیم، شامل مجموع درجه همسایگان نیز می‌شود و به صورت زیر تعریف می‌شود:

$$W_i^{D2} = W_i^{D1} + \sum_{j \in N_i} W_j^{D1}$$

این فرمول، اطلاعات بیشتری از گام دوم همسایگی را برای گره  $i$  اضافه می‌کند.

#### ۴-۲-۱-۴ ماتریس مرکزیت درجه‌ای همسایگی تعمیم‌یافته ( $W^{D3}$ )

در این ماتریس، اطلاعات گام سوم همسایگی نیز با استفاده از فرمول زیر لحاظ می‌شود:

$$W_i^{D3} = W_i^{D2} + \sum_{j \in N_i} W_j^{D2}$$

#### ۵-۲-۱-۴ ماتریس مرکزیت شاخص $H$ مستقیم همسایگی ( $W^H$ )

این ماتریس بر اساس شاخص  $H$  محاسبه شده و به صورت زیر تعریف می‌شود:

$$W_i^H = H_i (\text{Set}\{W_j^{D1} \mid j \in N_i\})$$

که در آن:

---

<sup>5</sup>Adjacency Matrix

• نشان دهنده شاخص  $H_i$  گره  $i$  است.

• مجموعه‌ای شامل درجه همسایگان گره  $i$  است.  $\text{Set}\{W_j^{D^1} \mid j \in N_i\}$

#### ۴-۱-۲-۶ ماتریس مرکزیت شاخص $H$ تجمعی همسایگی ( $W^{H^2}$ )

این ماتریس اطلاعات شاخص  $H$  همسایگان در گام دوم را نیز در نظر می‌گیرد و به صورت زیر تعریف می‌شود:

$$W_i^{H^2} = W_i^{H^1} + \sum_{j \in N_i} W_j^{H^1}$$

#### ۴-۱-۲-۷ ماتریس مرکزیت شاخص $H$ تعمیم‌یافته همسایگی ( $W^{H^3}$ )

این ماتریس اطلاعات گام سوم همسایگی را با استفاده از فرمول زیر اضافه می‌کند:

$$W_i^{H^3} = W_i^{H^2} + \sum_{j \in N_i} W_j^{H^2}$$

#### ۴-۱-۲-۸ ساخت ماتریس مجاورت

برای هر گره، یک ماتریس مجاورت با اندازه ثابت  $(L \times L)$  بر اساس همسایه‌های یک مرحله‌ای ساخته می‌شود. در صورتی که تعداد همسایه‌ها کمتر از  $L$  باشد، از لایه‌گذاری صفر<sup>۶</sup> برای پر کردن ماتریس استفاده می‌شود. در صورتی که تعداد همسایگان گره‌ای بیشتر از  $L$  باشد، به منظور محدود کردن اندازه‌ی ماتریس مجاورت به  $L \times L$ ، از استراتژی‌های زیر استفاده می‌شود:

۱. پس از اینکه مقادیر  $W^{D^3}$  برای تمامی همسایگان گره  $i$  محاسبه شد، همسایگان بر اساس مقادیر

به ترتیب نزولی مرتب می‌شوند:

$$N'(i) = \text{Sort}(N(i), \text{by } W^{D^3}, \text{descending})$$

<sup>6</sup>Zero-Padding

	datasets > Features > figeys.csv > data
	You, last month   1 author (You)
1	Node,WiD1,WiD2,WiD3,WiH1,WiH2,WiH3
2	1,27,280,10994,10,185,2601
3	2,14,1634,15987,12,186,10307
4	3,14,162,11314,8,139,2038
5	4,1,15,177,1,9,148
6	5,41,280,17961,10,220,3093 You, last month
7	6,12,1422,15537,11,173,10633
8	7,12,1237,12047,11,158,8396
9	8,110,1632,108232,16,913,18213
10	9,19,2159,22778,19,286,15614
11	10,172,977,58042,12,564,8238
12	11,12,1310,14513,12,164,8754
13	12,12,98,9744,6,83,1221
14	13,2,16,119,2,9,96
15	14,23,1435,17812,12,212,10748
16	15,1,24,1459,1,13,225
17	16,23,137,6736,7,106,1350
18	17,10,699,6427,10,135,4670
19	18,7,52,802,3,27,314
20	19,2,15,119,2,9,76
21	20,9,155,8682,8,101,1635
22	21,1,24,161,1,8,114
23	22,12,998,9360,11,149,5705
24	23,5,404,3092,5,55,1666
25	24,30,220,13677,9,163,2272
26	25,7,399,3289,7,79,1971
27	26,5,505,4360,5,56,2627
28	27,19,229,14253,8,163,2722

شکل ۴-۲: ویژگی های استخراج شده از گراف Figeys در فایل csv ذخیره شده

۲. از بین لیست مرتب شده، تنها  $L$  همسایه‌ی اول انتخاب می‌شوند:

$$N'(i) = \{j_1, j_2, \dots, j_L \mid j_k \in N(i) \text{ و } W_{j_1}^{D^3} \geq W_{j_2}^{D^3} \geq \dots \geq W_{j_L}^{D^3}\}$$

۹-۲-۱-۴ ترکیب کanal‌ها و ساخت مجموعه کanal‌های ساختاری بازنمایی گره‌ها

برای هر گره، دو «مجموعه کanal ساختاری بازنمایی گره» ساخته می‌شود:

۱. کanal مبتنی بر درجه: شامل مقیاس‌های مختلف درجه

۲. کanal مبتنی بر شاخص  $H$ : شامل مقیاس‌های مختلف شاخص  $H$

پس از محاسبه مقیاس‌های فوق، دو مجموعه کanal برای هر گره ایجاد می‌شود:

- کanal مبتنی بر درجه:

$$E^D(i) = \{E^{D\backslash}(i), E^{D\setminus}(i), E^{D\triangle}(i)\}$$

- کanal مبتنی بر شاخص  $H$ :

$$E^H(i) = \{E^{H\backslash}(i), E^{H\setminus}(i), E^{H\triangle}(i)\}$$

مقادیر کanal‌ها به شرح زیر تعریف می‌شوند:

$$E_i^{D_t} = \begin{cases} a_{lk} + W_i^{D_t} & \text{if } l = k, \\ a_{\setminus k} W_k^{D_t} & \text{else if } k \neq \setminus, \\ a_{l\setminus} W_l^{D_t} & \text{else if } l \neq \setminus, \\ a_{lk} & \text{else} \end{cases}$$

$$E_i^{H_t} = \begin{cases} a_{lk} + W_i^{H_t} & \text{if } l = k, \\ a_{\setminus k} W_k^{H_t} & \text{else if } k \neq \setminus, \\ a_{l\setminus} W_l^{H_t} & \text{else if } l \neq \setminus, \\ a_{lk} & \text{else} \end{cases}$$

۱۰-۲-۱-۴ پیش‌بینی نهایی

پس از آماده‌سازی کanal‌ها، این ویژگی‌ها به شبکه عصبی پیچشی تغذیه می‌شوند. مدل شامل لایه‌های پیچشی و لایه‌های چگال<sup>۵</sup> است که از تابع Leaky ReLU برای فعال‌سازی استفاده می‌کند. در نهایت، خروجی مدل یک مقدار پیش‌بینی شده برای تاثیرگذاری هر گره است.

<sup>7</sup>Dense Layers

### ۳-۱-۴ پارامترهای مدل

- اندازه کرنل (Kernel Size) : (2, 2)

- ماکریم پولینگ (MaxPooling) : (2, 2)

- تعداد اپوکها<sup>۸</sup> : 200

- بهینه‌ساز<sup>۹</sup> : Adam با نرخ یادگیری<sup>۱۰</sup>  $5 \times 10^{-4}$

- تابع هزینه<sup>۱۱</sup> : میانگین مربع خطای (MSE)

### ۲-۴ سامانه توسعه یافته

در این پژوهه، توسعه‌ی رابط کاربری برای مدیریت و تحلیل داده‌های خروجی شبکه‌های پیچیده مورد توجه قرار گرفته است. این رابط کاربری از طریق یک نرم‌افزار وب محور طراحی شده که قادر به نمایش، تحلیل و پردازش داده‌ها به شکلی کاربرپسند می‌باشد. به‌طور خاص، این رابط به گونه‌ای طراحی شده است که امکان مشاهده و تجزیه و تحلیل شبکه‌ها و گره‌های مختلف را برای کاربران فراهم کند.

### ۱-۲-۴ توسعه رابط کاربری مبتنی بر وب

برای ایجاد این رابط، از کتابخانه Streamlit استفاده شده است. این کتابخانه که به طور خاص برای توسعه برنامه‌های داده‌محور و تحلیلی طراحی شده است، امکان ایجاد صفحات تعاملی<sup>۱۲</sup> و پیش‌خوان<sup>۱۳</sup> های بصری را فراهم می‌کند. مزایای اصلی استفاده از Streamlit شامل سهولت در استفاده، سرعت در توسعه، امکان تعاملات زنده، پشتیبانی از نمودارها و گراف‌ها و یکپارچگی با کتابخانه‌های پایتون می‌باشد.

<sup>8</sup>Epoch

<sup>9</sup>Optimizer

<sup>10</sup>Learning Rate

<sup>11</sup>Loss Function

<sup>12</sup>Interactive

<sup>13</sup>Dashboard

## ۲-۲-۴ مزایای استفاده از کتابخانه Streamlit

دلیل انتخاب کتابخانه Streamlit در پروژه، همخوانی ویژگی های آن با نیاز های پروژه و البته مزایای زیاد آن در مقایسه با کاندید های دیگر بود:

- سهولت در استفاده
- سرعت در توسعه
- تعاملات زنده
- پشتیبانی از گرافها
- یکپارچگی با پایتون

## ۳-۲-۴ صفحات رابط کاربری

رابط کاربری این پروژه شامل بخش های مختلفی است که هر کدام به منظوری خاص طراحی شده اند:

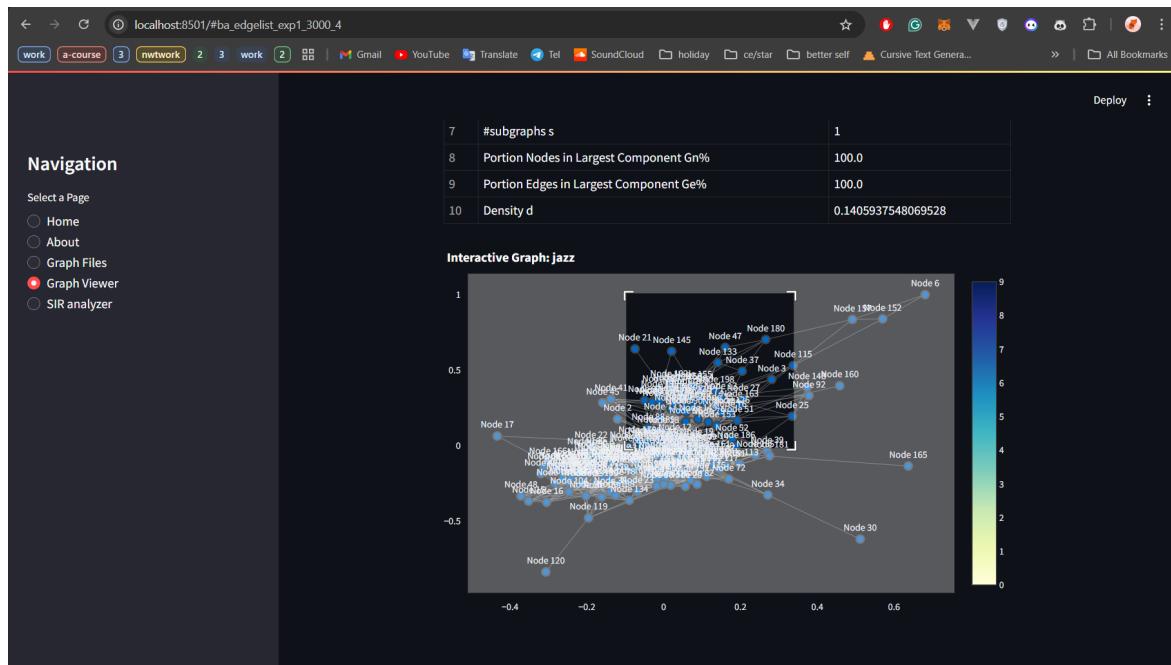
- صفحه مدیریت شبکه ها: این صفحه به کاربران اجازه می دهد تا شبکه های مختلف را بارگذاری، مشاهده و حذف کنند. در واقع، این صفحه امکان مدیریت مستقل داده های شبکه ای را فراهم می آورد.
- صفحه تحلیل و نمایش وضعیت برچسب ها: کاربران می توانند برچسب های تولید شده را مشاهده و تحلیل کنند. این بخش برای بررسی نتایج الگوریتم های مختلف رتبه بندی گره ها طراحی شده است.
- صفحه نمایش خروجی های شبکه: این بخش نتایج تحلیلی هر گراف را نمایش می دهد و اطلاعات مرتبط با گره ها و ویژگی های آنها را ارائه می کند.
- صفحه تحلیل گره های مهم: در این صفحه، گره های مهم شبکه بر اساس معیارهای رتبه بندی نمایش داده می شوند.
- صفحه تحلیل زمانی: این صفحه زمان های مختلف اجرای الگوریتم ها و تحلیل های مربوطه را نمایش می دهد و بهبود کارایی و انتخاب الگوریتم مناسب کمک می کند.

## ۴-۲-۴ پیش‌پردازش داده‌ها برای رابط کاربری

یکی از ویژگی‌های مهم این پروژه، توسعه رابط کاربری پویا و تعاملی برای نمایش و تحلیل شبکه‌های گراف پیچیده بود. این رابط کاربری به کاربران این امکان را می‌دهد که به صورت پویا اطلاعات گراف و حتی شکل گراف را مشاهده و با آن تعامل کنند تا درک بهتری از ساختار و ویژگی‌های شبکه‌ها به دست آورند. با این حال، همانطور که می‌دانیم، در شبکه‌های گراف پیچیده ممکن است تعداد گره‌ها به چند ده هزار گره و تعداد یال‌ها به چند صد هزار یال برسد. در چنین شرایطی، بارگذاری و نمایش آنی گراف‌ها در رابط کاربری به‌ویژه در صورت عدم پیش‌پردازش داده‌ها، می‌تواند باعث ایجاد زمان‌های تأخیر طولانی و غیرمنطقی شود که تجربه کاربری را به شدت کاهش می‌دهد. برای این منظور گراف‌ها بارگزاری شده، موقعیت گره‌ها بر اساس چیدمان فنری گراف محاسبه و داده‌های پردازش شده در قالب فایل‌های Pickle ذخیره شدند که شامل اطلاعات گراف، موقعیت گره‌ها، و مشخصات یال‌ها بودند. برای بهبود عملکرد رابط کاربری و کاهش زمان انتظار کاربران، فرآیند دیگری برای استخراج و ذخیره اطلاعات آماری گراف‌ها (مانند تعداد گره‌ها، تعداد یال‌ها، توزیع درجات گره‌ها و ضریب خوشبندی متوسط شبکه) پیاده‌سازی شد. هدف از این فرآیند، آماده‌سازی داده‌های پایه‌ای گراف‌ها به صورت پیش‌پردازش شده و ذخیره آن‌ها در قالب فایل‌های CSV بود. این‌گونه امکان دسترسی سریع و کارآمد به ویژگی‌های اصلی هر گراف را فراهم شد.

این رویکرد، علاوه بر کاهش چشمگیر زمان بارگذاری صفحات، کاربران را قادر می‌سازد تا بدون نیاز به محاسبات اضافی، تحلیل‌های اولیه خود را انجام دهند و اطلاعات گراف‌ها را به سرعت بررسی کنند. به‌ویژه در مواردی که حجم گراف‌ها بسیار بزرگ است، این فرآیند به عنوان یک راه حل کلیدی برای افزایش کارایی و بهبود تجربه کاربری در نظر گرفته شد.

تصاویر ۳-۴ و ۴-۴ نمایی از رابط کاربری تعاملی را نشان می‌دهند که کاربر را در تحلیل گراف‌ها و بررسی ویژگی‌های شبکه یاری می‌کند.



شکل ۴-۳: نمایش گراف‌ها در رابط کاربری تعاملی برای تحلیل شبکه‌ها



شکل ۴-۴: رابط کاربری تعاملی برای تحلیل مقیاس تأثیرگذاری گره‌ها با استفاده از شبیه‌سازی SIR

## فصل پنجم

### ارزیابی

## ۱-۵ مقدمه

در این بخش، عملکرد مدل‌ها و نتایج بدست‌آمده از ارزیابی آنها بررسی می‌شود. ابتدا به تعریف معیارهای ارزیابی می‌پردازیم. سپس تنظیمات آزمایش‌ها و پارامترهای مورد استفاده را توضیح می‌دهیم. در نهایت، نتایج تجربی به همراه تحلیل عملکرد مدل‌ها ارائه خواهد شد.

## ۲-۵ تعاریف معیارهای ارزیابی

در این قسمت، معیارهای ارزیابی مانند ضریب همبستگی کندال و میانگین دقت متوسط معرفی و توضیح داده می‌شوند تا ابزار مناسبی برای سنجش دقت و کیفیت رتبه‌بندی مدل فراهم گردد.

### ۱-۲-۵ ضریب همبستگی کندال (Kendall's $\tau$ Correlation Coefficient)

ضریب همبستگی کندال<sup>۱</sup> یک معیار آماری است که برای ارزیابی میزان شباهت یا توافق بین دو لیست رتبه‌بندی شده، مورد استفاده قرار می‌گیرد. این معیار بر اساس شمارش تعداد جفت‌های موافق<sup>۲</sup> و جفت‌های مخالف<sup>۳</sup> در دو لیست ساخته می‌شود. مقدار این ضریب عددی بین -۱ و ۱ است.

فرمول:

$$\tau(L_1, L_2) = \frac{2(N_c - N_d)}{n(n-1)} \quad (1-5)$$

در این فرمول:

- $L_1$  و  $L_2$ : دو لیست رتبه‌بندی شده که طول یکسان دارند

- $N_c$ : تعداد جفت‌های موافق بین دو لیست

- $N_d$ : تعداد جفت‌های مخالف بین دو لیست

<sup>1</sup>Kendall's  $\tau$

<sup>2</sup>Concordant Pairs

<sup>3</sup>Discordant Pairs

- $n$ : تعداد آیتم‌ها در هر لیست

تعریف جفت‌های موافق و مخالف:

- **جفت‌های موافق (Concordant Pairs)**: اگر برای دو آیتم  $i$  و  $j$ , رتبه‌بندی آن‌ها در هر دو لیست

به یک ترتیب باشد (یعنی اگر  $i$  در  $L_1$  قبل از  $j$  باشد، در  $L_2$  نیز همینطور باشد)، این جفت به عنوان موافق شناخته می‌شود.

- **جفت‌های مخالف (Discordant Pairs)**: اگر ترتیب رتبه‌بندی آیتم‌های  $i$  و  $j$  در  $L_1$  و  $L_2$  معکوس باشد، این جفت مخالف محسوب می‌شود.

ویژگی‌ها:

- اگر  $1 = \tau$ : لیست‌ها به طور کامل با یکدیگر موافق هستند.
- اگر  $0 = \tau$ : لیست‌ها هیچ رابطه مشخصی با یکدیگر ندارند.
- اگر  $1 - \tau$ : لیست‌ها به طور کامل مخالف یکدیگر هستند.

مزایا:

- به تغییرات کوچک در رتبه‌بندی حساس است.
- برای مقایسه رتبه‌بندی‌های نسبی بسیار مفید است.
- بر خلاف معیارهای همبستگی دیگر (مانند پیرسون<sup>4</sup>), برای داده‌های غیرخطی نیز مناسب است.

## Mean Average Precision (MAP) ۲-۲-۵

میانگین متوسط دقت (MAP) یکی از معیارهای محبوب در حوزه بازیابی اطلاعات و تحلیل دقت الگوریتم‌های یادگیری ماشین است. این معیار نشان‌دهنده میزان دقت و قابلیت اعتماد به الگوریتم در پیش‌بینی و رتبه‌بندی صحیح عناصر در یک مجموعه داده است. به عبارت ساده‌تر،

<sup>4</sup>Pearson

این معیار بررسی می‌کند که تا چه حد پیش‌بینی‌های الگوریتم با نتایج واقعی هم‌خوانی دارد و به طور خاص برای رتبه‌بندی آیتم‌ها مناسب است.

فرمول: فرمول کلی این معیار به صورت زیر تعریف می‌شود:

$$\text{MAP}(L_1, L_2) = \frac{1}{k} \sum_{i=1}^k \frac{c}{i} \cdot \text{rel}(i) \quad (2-5)$$

که در این فرمول:

- $L_1$  و  $L_2$ : دو لیست رتبه‌بندی شده هستند.  $L_1$  معمولاً لیست پیش‌بینی شده توسط الگوریتم و  $L_2$  لیست مرجع (لیست واقعی) است.
- تعداد کل آیتم‌های موجود در لیست  $L_1$ .
- تعداد پیش‌بینی‌های درست (مرتب) در میان اولین  $i$  آیتم‌های  $L_1$ .
- $i$ : شاخص (رتبه) هر آیتم در لیست رتبه‌بندی  $L_1$ .
- $c$ : تابعی نشانگر که مقدار آن برابر با ۱ است اگر آیتم در رتبه  $i$  مرتب باشد و در غیر این صورت مقدار ۰ خواهد بود.

عملکرد: MAP عملکرد الگوریتم را در تمامی رتبه‌ها ارزیابی می‌کند. به این صورت که هرچه آیتم‌های مرتب در رتبه‌های بالاتری قرار گیرند، مقدار MAP بیشتر خواهد بود. مقدار نهایی MAP بین ۰ و ۱ قرار دارد:

- مقدار نزدیک به ۱: نشان‌دهنده عملکرد بسیار خوب الگوریتم
- مقدار نزدیک به ۰: نشان‌دهنده ضعف الگوریتم در پیش‌بینی یا رتبه‌بندی

برای به دست آوردن MAP در شرایطی که چندین احتمال یا مجموعه داده مختلف وجود دارد (مانند چند مقدار احتمالی  $\beta$  در مدل انتشار)، میانگین تمام مقادیر MAP محاسبه می‌شود:

$$\text{AvgMAP}(\{L_{\alpha_{\min}}, \dots, L_{\alpha_{\max}}\}, L_1) = \frac{1}{m} \sum_{\alpha=\alpha_{\min}}^{\alpha_{\max}} \text{MAP}(L_\alpha, L_1) \quad (3-5)$$

که در این فرمول:

- $m$ : تعداد کل حالات احتمالی.
- $L_\alpha$ : لیست‌های رتبه‌بندی شده برای مقادیر مختلف احتمالی  $\alpha$ .
- کاربردها: MAP به ویژه در مسائل زیر استفاده می‌شود:
- بازیابی اطلاعات (Information Retrieval): ارزیابی کیفیت سیستم‌های جستجو برای بازگرداندن اسناد مرتبط.
- تشخیص اشیاء (Object Detection): ارزیابی دقت الگوریتم‌های یادگیری عمیق در شناسایی و طبقه‌بندی اشیاء.
- تحلیل شبکه‌های پیچیده: بررسی توانایی الگوریتم‌ها در شناسایی گره‌های تأثیرگذار، مانند این پروژه.

### ۳-۲-۵ ضرورت معیارهای ارزیابی انتخاب شده در ارزیابی

معیارهای مذکور برای ارزیابی مدل‌های آموزش‌داده شده استفاده می‌شود. انتخاب این معیارها به دلایل زیر ضروری بوده و اهداف پژوهش را به خوبی پوشش می‌دهد.

#### ۱-۳-۲-۵ توانایی ارزیابی دقیق رتبه‌بندی‌ها

مدل معرفی شده، به منظور شناسایی و رتبه‌بندی گره‌های تأثیرگذار در شبکه‌های پیچیده طراحی شده است. هر یک از معیارهای انتخاب شده از جنبه‌های مختلف دقت و کیفیت رتبه‌بندی را ارزیابی می‌کند:

- ضریب همبستگی کندال: بررسی دقیق روابط نسبی بین گره‌ها و تحلیل تغییرات کوچک در رتبه‌بندی را ممکن می‌سازد.
- ضریب همبستگی رتبه‌ای اسپیرمن: این معیار، همبستگی کلی بین رتبه‌های پیش‌بینی شده و رتبه‌های واقعی را اندازه‌گیری کرده و اثر تغییرات یکنواخت در رتبه‌ها را بررسی می‌کند.
- میانگین متوسط دقت: این معیار نشان‌دهنده توانایی مدل در رتبه‌بندی صحیح گره‌های مرتبط در صدر لیست است و به ویژه در شناسایی گره‌های با اهمیت بالا اهمیت دارد.

### ۲-۳-۲-۵ پوشش جنبه‌های مختلف دقت در شبکه‌های پیچیده

شبکه‌های پیچیده به دلیل داشتن ساختارهای غیرخطی و پویا، نیازمند معیارهایی هستند که بتوانند ویژگی‌های متنوع رتبه‌بندی را تحلیل کنند:

- ضریب همبستگی کندال: برای تحلیل روابط محلی و رتبه‌بندی‌های جفتی طراحی شده است و تغییرات جزئی در رتبه‌ها را به خوبی منعکس می‌کند.
- ضریب همبستگی رتبه‌ای اسپیرمن: می‌تواند به بررسی روابط کلی‌تر و تغییرات یکنواخت رتبه‌ها بپردازد.
- میانگین متوسط دقت: با تمرکز بر آیتم‌های صدر لیست، به تحلیل رتبه‌بندی آیتم‌های مهم در شبکه می‌پردازد.

### ۳-۳-۲-۵ هماهنگی با اهداف مدل

این مدل بر شناسایی گره‌های تأثیرگذار با استفاده از ترکیب اطلاعات محلی و ویژگی‌های ساختاری شبکه مرکز است. معیارهای انتخاب شده به طور خاص برای این اهداف مناسب هستند:

- ضریب همبستگی کندال و ضریب همبستگی رتبه‌ای اسپیرمن: با ارزیابی میزان تطابق رتبه‌بندی‌های مدل با داده‌های واقعی، کارایی مدل در استخراج ویژگی‌های کلیدی را می‌سنجد.
- میانگین متوسط دقت توانایی مدل را در شناسایی گره‌های تأثیرگذار با اهمیت بالا به صورت کمی تحلیل می‌کند.

### ۴-۳-۲-۵ سازگاری با تحلیل عملکرد در مقیاس‌های مختلف

معیارهای انتخاب شده در ارزیابی شبکه‌های مختلف و مقیاس‌های متنوع عملکرد دارند:

- ضریب همبستگی کندال و ضریب همبستگی رتبه‌ای اسپیرمن: امکان تحلیل دقیق شبکه‌های کوچک و بزرگ با تنوع بالا در ساختار را فراهم می‌کنند.
- میانگین متوسط دقت: برای شبکه‌های پیچیده و بزرگ که نیاز به رتبه‌بندی گره‌ها با دقت بالا دارند، بسیار کاربردی است.

### ۵-۲-۳ کاهش پیچیدگی محاسباتی

در حالی که معیارهایی مانند مرکزیت بینایی و مرکزیت نزدیکی ممکن است به دلیل پیچیدگی محاسباتی زیاد برای ارزیابی شبکه‌های بزرگ نامناسب باشند، معیارهای انتخاب شده در این پژوهش، تعادل بین دقت و هزینه محاسباتی را فراهم می‌کنند:

- ضریب همبستگی کنдал و ضریب همبستگی رتبه‌ای اسپیرمن: محاسبات ساده‌ای دارند و به سرعت قابل اجرا هستند.
- میانگین متوسط دقت: با تمرکز بر رتبه‌های بالا، پیچیدگی محاسبات را کاهش داده و نتایج معناداری ارائه می‌دهد.

### ۶-۲-۳ مقایسه و تحلیل جامع عملکرد مدل

انتخاب ترکیبی از این سه معیار، تحلیل چندجانبه‌ای از عملکرد مدل فراهم می‌کند:

- ضریب همبستگی کنдал و ضریب همبستگی رتبه‌ای اسپیرمن: برای مقایسه و تحلیل دقیق همبستگی رتبه‌بندی‌ها استفاده می‌شوند.
- میانگین متوسط دقت: مکمل این معیارها بوده و به تحلیل دقت رتبه‌بندی‌های صدر لیست می‌پردازد.

## ۳-۵ تنظیمات اولیه و پیکربندی آزمایش‌ها و پارامترها

### ۱-۳-۵ تنظیمات بهینه‌سازی و پارامترهای آموزش

برای آموزش مدل، پارامترهای مختلفی تنظیم شدند تا بهینه‌ترین شرایط برای یادگیری فراهم شود. این پارامترها عبارتند از:

- نرخ یادگیری (Learning Rate): مقدار 0.0005 انتخاب شده است که تعادل بین سرعت همگرایی و دقت نهایی مدل را حفظ می‌کند.
- تعداد Epochs: ابتدا تعداد 200 و پس از بررسی تاثیر کاهش تعداد epoch و اطمینان از یادگیری کامل مدل، تعداد epoch 150 تنظیم شد.

- اندازه Batch: مقدار 4 برای batch size انتخاب شده است تا از توازن بین استفاده بهینه از حافظه و دقت مدل اطمینان حاصل شود.
- تابع خطا (Loss Function): از خطای میانگین مربعات (MSE) برای ارزیابی تفاوت بین مقادیر پیش‌بینی‌شده و مقادیر واقعی استفاده شد.
- بهینه‌ساز (Optimizer): بهینه‌ساز Adam برای بهبود سرعت همگرایی و جلوگیری از گیر افتادن در مینیمم محلی به کار رفت.

## ۲-۳-۵ مشخصات فنی اجرای آزمایش‌ها

- آزمایش‌ها بر روی یک سیستم مجهرز به سخت‌افزار و نرم‌افزار زیر انجام شدند:
- کارت گرافیک (GPU): NVIDIA GeForce MX150 با حافظه گرافیکی 2.15 GB
  - پردازنده مرکزی (CPU): سیستم به پردازنده Intel64 Family 6 Model 142 Stepping 11، GenuineIntel مجهرز بوده که شامل 4 هسته فیزیکی و 8 هسته منطقی است. فرکانس پردازنده تا 1992.0 MHz متغیر بوده و در زمان آزمایش‌ها میزان استفاده از پردازنده 86.6% گزارش شده است.
  - سیستم‌عامل: سیستم‌عامل Windows 10 با نسخه 10.0.19045 روی دستگاه اجرا می‌شد. و معماری آن AMD64 است.
  - کتابخانه‌ها و ابزارها: کتابخانه‌های PyTorch برای پیاده‌سازی شبکه عصبی، NetworkX برای تحلیل شبکه، و Matplotlib برای رسم نمودارها استفاده شدند.

## ۴-۵ تحلیل عملکرد

### ۱-۴-۵ دقت مدل در شناسایی گره‌های برتر

مدل پیشنهادی برای پیش‌بینی مقادیر SIR و شناسایی گره‌های تأثیرگذار در شبکه‌های پیچیده طراحی شده است. به منظور تحلیل دقیق عملکرد مدل در این زمینه، ابتدا به بررسی بسیار کوتاه توزیع مقادیر SIR در گراف‌های مختلف پرداخته شده است. این بررسی نشان می‌دهد که گراف‌های مورد استفاده

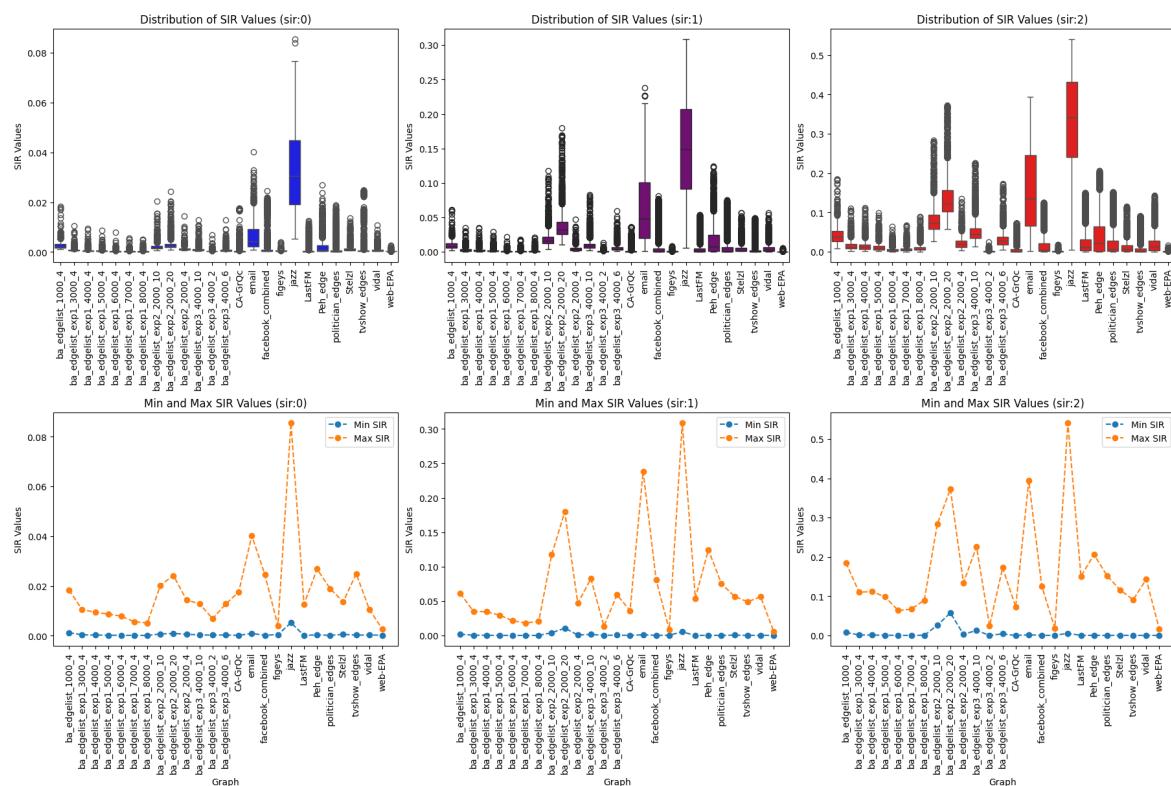
دارای ساختارهای متنوعی هستند و مقادیر SIR در میان گرافها دارای توزیع‌های متفاوتی است. سپس عملکرد مدل در پیش‌بینی دقیق این مقادیر و رتبه‌بندی گره‌ها تحلیل شده است.

#### ۱-۱-۴-۵ بروزی توزیع مقادیر SIR

برای ارزیابی دقیق عملکرد مدل در پیش‌بینی مقادیر SIR، ابتدا توزیع این مقادیر در گراف‌های مختلف بررسی شده است. شکل ۱-۵ دو نمودار را نشان می‌دهد:

- نمودار Boxplot: این نمودار توزیع مقادیر SIR در گراف‌های مختلف را نمایش می‌دهد و نشان‌دهنده وجود اختلافات ساختاری در گراف‌ها است.

- نمودار Min-Max: این نمودار مقادیر حداقل و حداکثر SIR در هر گراف را نشان می‌دهد که بیان‌گر دامنه مقادیر برای هر شبکه است.



شکل ۱-۵: توزیع مقادیر SIR در گراف‌های مختلف (نمودار Boxplot و Min-Max)

این نتایج نشان می‌دهد که گراف‌های مورد بررسی دارای ساختارهای متنوعی هستند. عملکرد خوب مدل پیشنهادی در این شرایط بیان‌گر توانایی آن در تعمیم‌پذیری<sup>۵</sup> و جلوگیری از بیش‌برازش<sup>۶</sup> است.

<sup>5</sup>Generalization

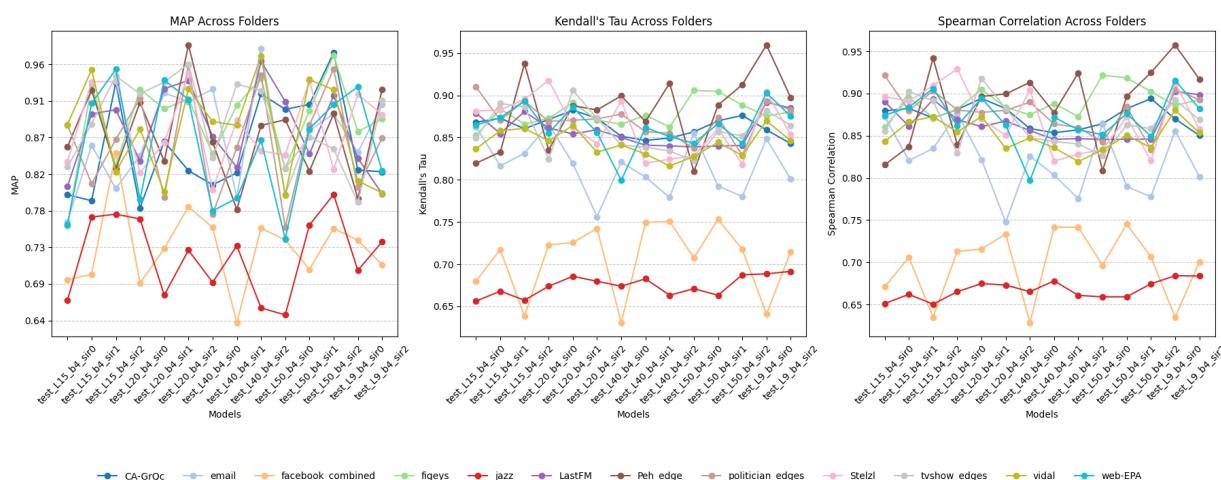
<sup>6</sup>Overfitting

## ۲-۴-۵ مقایسه عملکرد مدل‌های پیشنهادی

برای تحلیل دقیق‌تر، عملکرد مدل در دو دسته از گراف‌ها بررسی شده است:

- گراف‌های مصنوعی: همانطور که در قسمت تعاریف فصل داده‌ها ذکر شد، این گراف‌ها بر اساس مدل مقیاس آزاد طراحی شده‌اند و دارای ساختارهای یکنواخت‌تری هستند.
- گراف‌های دنیای واقعی: شامل شبکه‌های اجتماعی، زیستی و فناوری که دارای ساختارهای پیچیده‌تر و متنوع‌تری هستند.

شکل ۲-۵ و ۳-۵ عملکرد مدل را در هر دو دسته نشان می‌دهد. معیارهای MAP، Kendall's Tau و Spearman Correlation برای ارزیابی دقت مدل استفاده شده‌اند.



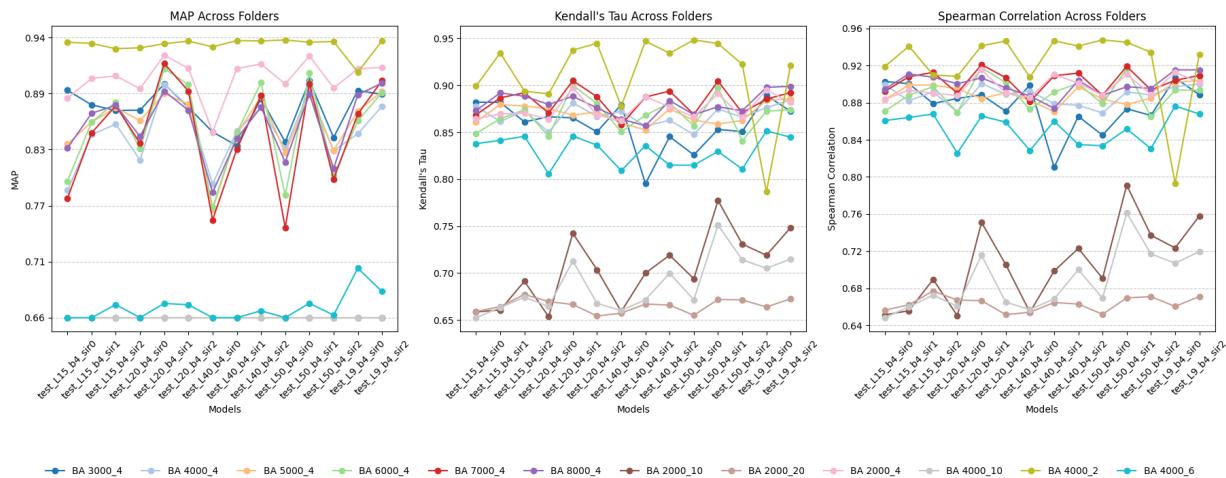
شکل ۲-۵: عملکرد مدل در گراف‌های دنیای واقعی بر اساس معیارهای مختلف

نتایج نشان می‌دهد که مدل پیشنهادی در هر دو نوع گراف عملکرد بسیار خوبی داشته و توانسته است گره‌های تأثیرگذار را به درستی شناسایی کند.

نکته‌ای که شاید بهتر باشد به آن دقت کنیم این است که در هر دو شکل، برای شبکه‌های مصنوعی یا شبکه‌های دنیای واقعی، میتوان اکثراً افزایش دقت را زمانی دید که مقدار آلفا (sir\_alpha) برابر با ۱.۵ است. این در حالی است که برای مقدار آلفای ۱ یا ۲ دقت کاهش می‌یابد. همانطور که در فصل داده‌ها ذکر شد (۲-۲-۳)، این بدلیل ویژگی‌های شبیه‌سازی SIR است.

• sir\_alpha = 1: احتمال انتقال بسیار پایین بوده و اثر توپولوژی شبکه دیده نمی‌شود.

• sir\_alpha = 2: احتمال انتقال بسیار بالا بوده و بیماری مستقل از تأثیر گره‌ها گسترش می‌یابد.

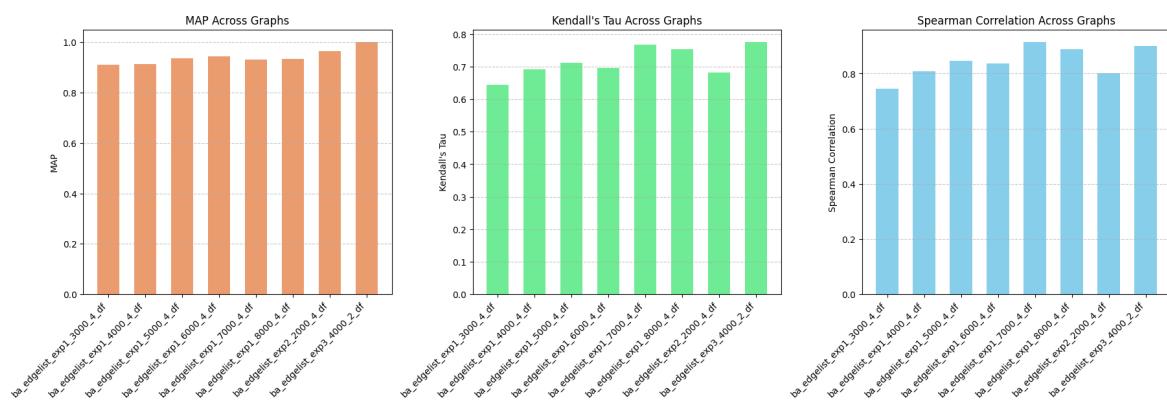


شکل ۳-۵: عملکرد مدل در گرافهای مصنوعی و بر اساس معیارهای مختلف

مقدار بهینه که تعادل مناسبی برای مشاهده تأثیر توپولوژی شبکه ایجاد می‌کند.

#### ۱-۲-۴-۵ تحلیل دقت مدل برای یک تنظیم خاص

به منظور تحلیل جزئی‌تر، عملکرد مدل در یک تنظیم خاص (test\_L15\_b4\_sir2) مورد بررسی قرار گرفته است. شکل ۴-۵ دقت مدل را در پیش‌بینی مقادیر SIR و رتبه‌بندی گره‌های برتر برای گرافهای Spearman Correlation و Kendall's Tau، معیارهای MAP و Kendall's Tau نمایش می‌دهد. در این تحلیل، معیارهای MAP، Kendall's Tau و Spearman Correlation مختلف نمایش می‌دهد. استفاده شده‌اند.



شکل ۴-۵: دقت مدل در معیارهای مختلف برای گرافهای مختلف در تنظیم 2

این تحلیل نشان می‌دهد که مدل در بیشتر گراف‌ها عملکرد پایداری داشته و توانسته است گره‌های تأثیرگذار را با دقت بالا شناسایی کند. اختلافات جزئی در برخی گراف‌ها نشان‌دهنده تأثیر ساختار شبکه بر دقت مدل است، اما میانگین عملکرد همچنان در سطح بالایی باقی مانده است.

## ۲-۴-۵ جمع‌بندی

تحلیل‌های انجام شده نشان می‌دهد که مدل پیشنهادی توانسته است در گراف‌های با ساختارهای متنوع و معیارهای مختلف، دقت بالایی ارائه دهد. این عملکرد برجسته نشان‌دهنده کارایی مدل در شناسایی گره‌های تأثیرگذار و پتانسیل آن برای کاربردهای گسترده در تحلیل شبکه‌های پیچیده است. همچنین این نکته را نشان میدهد که در آینده بهتر است در شبیه‌سازی SIR از مقادیر میانی SIR\_alpha در بازه‌ی [1, 9.1] استفاده کرد.

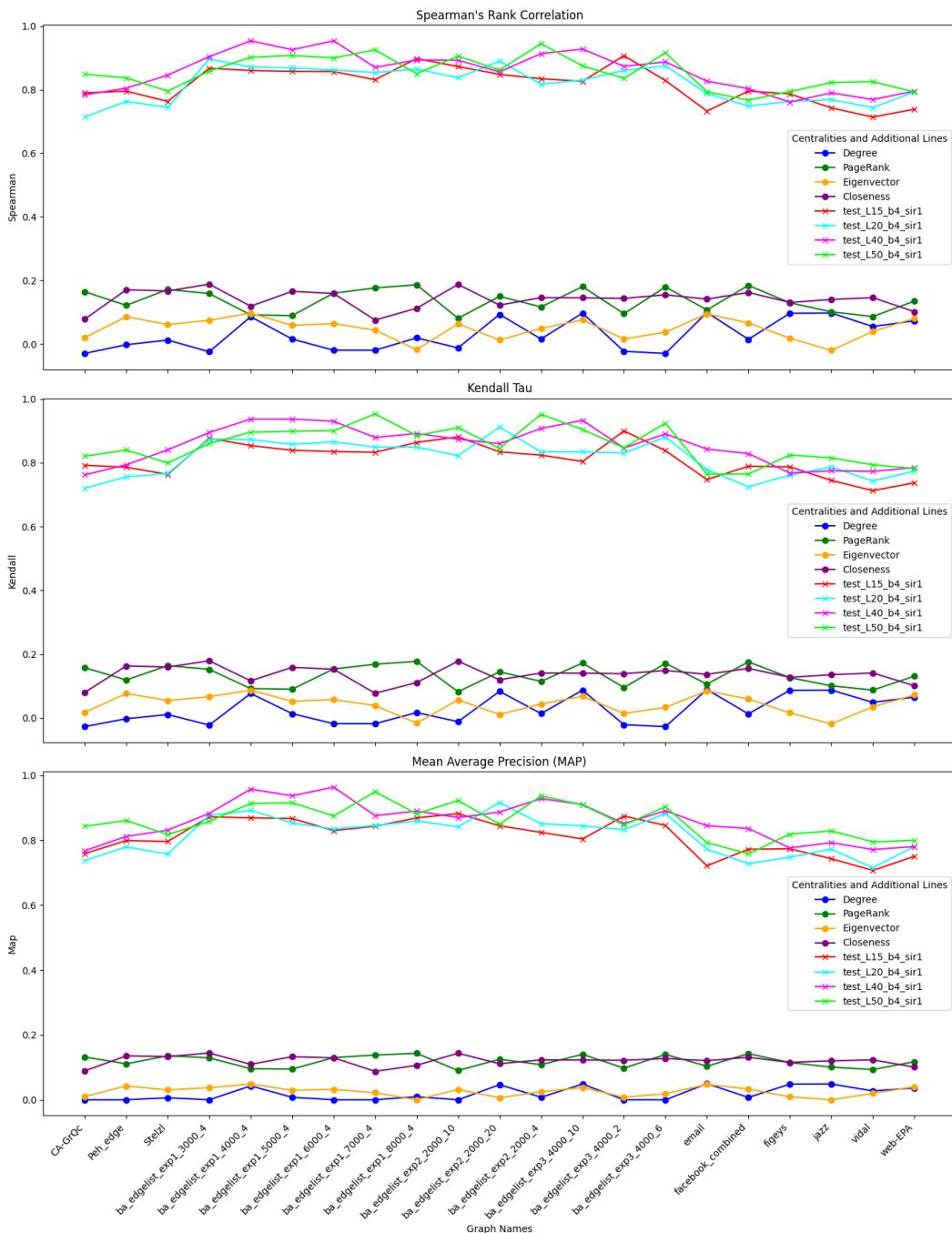
## ۳-۴-۵ معرفی الگوریتم‌های پایه

الگوریتم‌های مرکزیت پایه‌ای که در این پژوهش استفاده شده در این قسمت به اختصار توضیح داده شده‌اند. این الگوریتم‌ها شامل چهار معیار مرکزیت است که به منظور مقایسه عملکرد مدل پیشنهادی مورد استفاده قرار گرفته‌اند:

۱. مرکزیت نزدیکی (**Closeness Centrality**): فاصله میان یک گره و سایر گره‌ها را در یک شبکه محاسبه می‌کند و نشان می‌دهد که یک گره چقدر سریع می‌تواند اطلاعات را به دیگر گره‌ها منتقل کند. اگرچه همانطور که خواهیم دید این روش نیز دقت زیادی ندارد. در واقع، این معیار با در نظر گرفتن فاصله میان گره‌ها، تأثیر توپولوژی شبکه را بررسی می‌کند.
۲. مرکزیت رتبه صفحه (**PageRank**): برای شناسایی گره‌هایی که به گره‌های مهم دیگر متصل هستند، طراحی شده است. درواقع، این الگوریتم از احتمال انتقال اطلاعات در طول یال‌های شبکه استفاده می‌کند تا اهمیت هر گره را تعیین کند.
۳. مرکزیت بردار ویژه (**Eigenvector Centrality**): میزان تأثیرگذاری یک گره را بر اساس اتصال آن به گره‌های تأثیرگذار دیگر ارزیابی می‌کند. مرکزیت بردار ویژه به گره‌هایی که به سایر گره‌های پرنفوذ متصل هستند، وزن بیشتری می‌دهد و تأثیرگذاری آنها را تقویت می‌کند.
۴. مرکزیت درجه (**Degree Centrality**): تعداد گره‌های متصل به یک گره را اندازه‌گیری می‌کند. مرکزیت درجه ساده‌ترین روش برای ارزیابی اهمیت گره است و میزان تعامل مستقیم آن را با دیگر گره‌ها نشان می‌دهد.

### ۱-۳-۴-۵ مقایسه با الگوریتم‌های پایه

در این بخش، نتایج مدل پیشنهادی با الگوریتم‌های پایه مقایسه شده است. همان‌طور که در شکل ۵ نشان داده شده است، معیارهای ارزیابی شامل Kendall's τ و MAP بوده‌اند. در این شکل، مقادیر مربوط به گراف‌هایی که مقدار  $\text{alpha}_{\text{sir}}$  برابر با ۱ دارند، به دلیل اختصار نمایش داده شده است.



شکل ۵: مقایسه عملکرد مدل‌های پیشنهادی با الگوریتم‌های پایه در معیارهای مختلف.

بر اساس نتایج به دست آمده:

- الگوریتم‌های پایه مانند مرکزیت درجه‌ای و مرکزیت بردار ویژه در مقایسه با مدل پیشنهادی عملکرد ضعیفی نشان داده‌اند. مقدار Kendall's  $\tau$  و Spearman برای این الگوریتم‌ها به طور معمول زیر 0.05 بوده است.
- الگوریتم مرکزیت نزدیکی و مرکزیت رتبه صفحه‌ای نیز در معیارهای مختلف عملکرد متوسطی داشته، اما همچنان کمتر از مدل پیشنهادی.
- مدل پیشنهادی ما، مقادیر بالایی (بین 0.7 تا 1) در هر سه معیار نشان داده است که نشان‌دهنده دقیق و توانایی آن در شناسایی گره‌های تأثیرگذار است.

#### ۴-۴-۵ مزایا و نقاط قوت مدل پیشنهادی

مدل پیشنهادی در این پژوهه با بهره‌گیری از شبکه عصبی پیچشی گرافی و طراحی هوشمندانه ویژگی‌ها، توانسته است عملکرد برجسته‌ای در شناسایی و رتبه‌بندی گره‌های تأثیرگذار در شبکه‌های پیچیده ارائه دهد. مزایا و نقاط قوت این مدل به شرح زیر است:

##### ۱-۴-۴-۵ کارایی بالا در شناسایی گره‌های تأثیرگذار

مدل پیشنهادی با استفاده از ویژگی‌های ساختاری و محلی شبکه توانسته است گره‌های تأثیرگذار را با دقیق بالایی شناسایی کند. معیارهای ارزیابی مانند Kendall's  $\tau$ ، Spearman Rank Correlation و MAP و شناسانده‌نده کارایی این مدل در شناسایی و رتبه‌بندی گره‌ها هستند.

##### ۲-۴-۴-۵ کاهش پیچیدگی محاسباتی

در طراحی این مدل، از ویژگی‌های محلی ساده‌تری مانند درجه گره ( $W^D$ ) و شاخص  $H$  ( $W^H$ ) استفاده شده است. این ویژگی‌ها با وجود محاسبات ساده، اطلاعات کافی برای تحلیل ساختار شبکه را فراهم می‌کنند و موجب کاهش پیچیدگی محاسباتی مدل در مقایسه با روش‌های مبتنی بر ویژگی‌های پیچیده‌تر می‌شوند.

##### ۳-۴-۴-۵ انعطاف‌پذیری در مقیاس‌های مختلف شبکه

مدل پیشنهادی توانایی تحلیل شبکه‌هایی با اندازه‌ها و ساختارهای مختلف را دارد. استفاده از ماتریس‌های مجاورت با اندازه ثابت ( $L \times L$ ) و تنظیمات هوشمندانه در استخراج ویژگی‌ها، مدل را قادر ساخته است

تا بدون افت عملکرد در شبکه‌های بزرگ نیز کارایی خود را حفظ کند.

#### ۴-۴-۴-۵ تطبیق‌پذیری با انواع شبکه‌ها

مدل با استفاده از ماتریس‌های بازنمایی مبتنی بر کانال‌های ساختاری، قابلیت تحلیل شبکه‌های مختلف اجتماعی، زیستی، و فناوری را دارد. این تطبیق‌پذیری، مدل را به یک ابزار چندمنظوره برای تحلیل شبکه‌های پیچیده تبدیل کرده است.

#### ۴-۴-۵-۵ زمان آموزش و اجرا بهینه

با توجه به استفاده از ویژگی‌های محلی و طراحی بهینه شبکه عصبی، زمان موردنیاز برای آموزش و اجرای مدل در مقایسه با الگوریتم‌های پیچیده‌تر، بهینه‌سازی شده است. این ویژگی، مدل را برای کاربردهای واقعی و شبکه‌های با مقیاس بزرگ مناسب می‌سازد.

#### ۴-۴-۶-۵ قابلیت تفسیرپذیری

ویژگی‌های مورد استفاده در این مدل مانند درجه گره و شاخص  $H$ ، به راحتی قابل تفسیر هستند و می‌توانند به درک بهتری از رفتار گره‌های تأثیرگذار و ساختار شبکه کمک کنند. این قابلیت، مدل را برای تحقیقات علمی و کاربردهای عملی ارزشمندتر می‌کند.

#### ۷-۴-۴-۵ عملکرد برتر نسبت به الگوریتم‌های پایه

نتایج مقایسه‌ای مدل پیشنهادی با الگوریتم‌های پایه مانند مرکزیت درجه و مرکزیت بینابینی نشان داده است که این مدل توانسته دقیق‌تر و شناسایی‌تر از مدل پیشنهادی باشد.

#### ۴-۴-۸-۵ تطبیق با نیازهای کاربردی

مدل پیشنهادی با طراحی مازولار و استفاده از پارامترهای قابل تنظیم مانند  $L$  و تعداد epochs، امکان تنظیم دقیق برای پاسخگویی به نیازهای مختلف در کاربردهای شبکه‌ای را فراهم کرده است. این مزایا، مدل پیشنهادی را به یک ابزار قدرتمند برای تحلیل شبکه‌های پیچیده و شناسایی گره‌های تأثیرگذار تبدیل کرده و کارایی آن را در زمینه‌های مختلف به اثبات رسانده است.

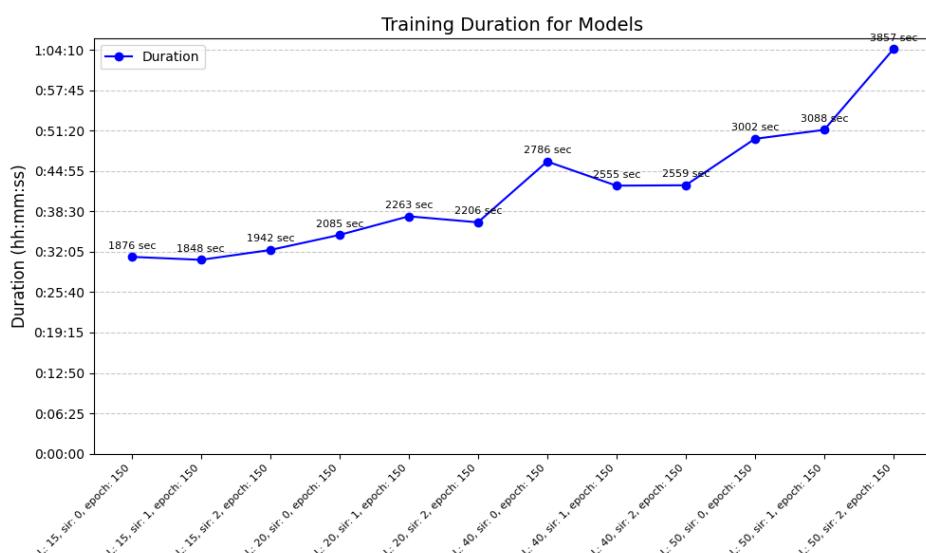
## ۴-۵ تحلیل مدت زمان آموزش و اجرا

مدت زمان آموزش و اجرای مدل‌های یادگیری ماشین یکی از معیارهای مهم در ارزیابی کارایی و قابلیت استفاده آنها در پروژه‌های واقعی است. در این بخش، تحلیل کاملی از زمان‌های آموزش و تست مدل ارائه شده است. این تحلیل شامل بررسی زمان‌های موردنیاز برای آموزش مدل‌ها در شرایط مختلف و مقایسه مدت زمان اعتبارسنجی آن‌ها در داده‌های آزمایشی می‌باشد.

### ۱-۵-۴ مدت زمان آموزش مدل‌ها

برای تحلیل مدت زمان آموزش مدل‌ها، مقادیر ثبت شده برای پارامترهای مختلف مانند تعداد اپوک‌ها (epochs)، اندازه ماتریس مجاورت ( $L$ )، و مقدار آلفا (sir\_alpha) بررسی شده‌اند. شکل ۵ نشان‌دهنده مدت زمان آموزش مدل‌ها با تنظیمات مختلف است.

- رابطه زمان آموزش با تعداد اپوک‌ها: با افزایش تعداد اپوک‌ها، مدت زمان آموزش به صورت خطی افزایش یافته است.
- اثر پارامتر  $L$ : مقادیر بزرگ‌تر  $L$  منجر به افزایش پیچیدگی محاسباتی و به تبع آن افزایش زمان آموزش شده‌اند.
- مقدار sir\_alpha: تغییرات این مقدار الگوی تأثیری ثابتی ندارد.

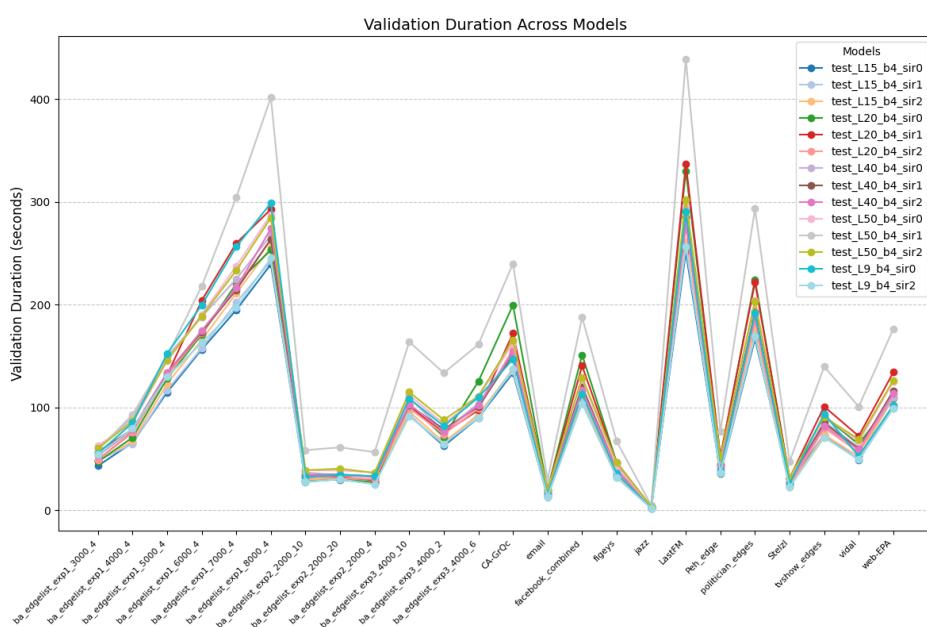


شکل ۵: مدت زمان آموزش مدل‌ها برای مقادیر مختلف پارامترهای  $L$ ، sir\_alpha، و epochs

## ۴-۵-۲ مدت زمان تست و اعتبارسنجی مدل‌ها

مدت زمان اعتبارسنجی مدل‌ها بر روی مجموعه‌ای از داده‌های آزمایشی با استفاده از گراف‌های مختلف محاسبه شده است. این تحلیل به بررسی زمان موردنیاز برای پردازش گراف‌ها توسط مدل‌های مختلف می‌پردازد. شکل ۷-۵ نشان‌دهنده مدت زمان تست مدل‌ها در گراف‌های آزمایشی است.

- مقایسه مدل‌ها: مدل‌های با تنظیمات پیچیده‌تر و مقدار  $L$  بزرگ‌تر زمان تست طولانی‌تری دارند.
- رابطه گراف و زمان تست: گراف‌های با تعداد گره‌ها و یال‌های بیشتر به‌طور طبیعی زمان بیشتری برای پردازش نیاز دارند.



شکل ۷-۵: مدت زمان تست مدل‌ها بر روی گراف‌های مختلف

## ۴-۵-۳ تحلیل کلی

تحلیل زمان اجرا نشان می‌دهد که مدل‌های توسعه‌یافته از نظر مدت زمان آموزش و تست عملکرد مناسبی دارند. با این حال، افزایش پارامترهایی مانند  $L$  منجر به افزایش زمان اجرا می‌شود، که باید در کاربردهای واقعی به دقت مدیریت شود. این تحلیل به کاربران کمک می‌کند تا با توجه به محدودیت‌های زمانی و منابع، تنظیمات بهینه‌ای برای مدل انتخاب کنند.

## **فصل ششم**

### **جمع‌بندی و نتیجه‌گیری و پیشنهادات**

## ۱-۶ جمع‌بندی و نتیجه‌گیری

چارچوب پیشنهادی این پژوهش به مسئله شناسایی گره‌های تأثیرگذار در شبکه‌های پیچیده می‌پردازد. اهمیت این موضوع در کاربردهای مختلفی مانند بازاریابی ویروسی، کنترل شایعات(بیماری با اطلاعات)، و مدیریت سیستم‌های اجتماعی و فناوری نهفته است. استفاده از شبکه‌های عصبی پیچشی و نمایش‌های محلی گره‌ها، باعث شده است که این رویکرد بتواند خصوصیات کلیدی گره‌ها را به شکلی کارآمد و دقیق شناسایی کند.

این روش با تمرکز بر استفاده از ماتریس مجاورت یک مرحله‌ای و بهره‌گیری از شاخص‌های ساده و کارآمدی مانند مرکزیت درجه و شاخص  $H$ ، تلاش کرده است تا هزینه‌های محاسباتی را کاهش دهد و نمایش منصفانه‌تری از گره‌ها ارائه دهد. برخلاف روش‌های سنتی که ممکن است در نمایش برخی ویژگی‌های شبکه ناکام باشند، این سیستم توانسته است با استفاده از چند مقیاس مختلف از شاخص‌ها، بینش عمیق‌تری از ساختار شبکه‌ها به دست آورد.

عملکرد این چارچوب در آزمایش‌های گسترده بر روی شبکه‌های واقعی و مصنوعی به خوبی نشان‌دهنده توانایی بالای آن بوده است.

یکی از ویژگی‌های برجسته‌ی آن، مقاومت آن در برابر تغییرات ساختاری شبکه‌ها است. برخلاف برخی الگوریتم‌های دیگر، این مدل توانسته است در مواجهه با تغییرات اندازه شبکه یا میانگین درجه گره‌ها عملکرد پایداری از خود نشان دهد. این ویژگی، آن را برای کاربردهایی که نیاز به انعطاف‌پذیری بالا دارند، به یک ابزار ارزشمند تبدیل کرده است.

آزمایش‌های انجام‌شده با استفاده از معیارهای دقیق ارزیابی، مانند ضریب همبستگی کندال و میانگین متوسط دقت، تأیید کرده‌اند که این سیستم به طور قابل توجهی در رتبه‌بندی گره‌ها بهتر از روش‌های دیگر عمل می‌کند. این آزمایش‌ها بر روی شبکه‌های متنوعی انجام شده‌اند که شامل انواع شبکه‌های اجتماعی، همکاری‌های پژوهشی، و شبکه‌های زیستی بوده‌اند. این نتایج نشان‌دهنده قدرت تعمیم‌پذیری بالای آن در مواجهه با ساختارهای مختلف شبکه است.

در مقایسه با روش‌های سنتی، این مدل توانسته است با رویکردی نوآورانه که از نمایش‌های محلی و تحلیل چند مقیاسی استفاده می‌کند، محدودیت‌های مربوط به شناسایی گره‌های کلیدی را کاهش دهد. استفاده از این روش، علاوه بر کاهش پیچیدگی زمانی، باعث بهبود دقت در شناسایی گره‌های تأثیرگذار شده است. این دستاوردها تأکید بیشتری بر پتانسیل روش‌های یادگیری عمیق در تحلیل شبکه‌های پیچیده دارد.

در مجموع، این پژوهه تلاش داشته است گام مهمی در جهت حل چالش‌های مربوط به شناسایی گره‌های تأثیرگذار بردارد و زمینه‌ساز تحقیقات بیشتر در این حوزه شود. نتایج به دست آمده نه تنها نشان‌دهنده کارایی بالای این مدل در تحلیل شبکه‌ها هستند، بلکه پتانسیل‌های یادگیری عمیق را برای حل مسائل پیچیده در دنیای واقعی برجسته می‌کنند. این پژوهش نقطه آغازی برای استفاده گسترده‌تر از روش‌های هوش مصنوعی در تحلیل شبکه‌های اجتماعی و علمی است.

## ۲-۶ پیشنهادات

این پژوهش تنها بر شناسایی گره‌های فردی متمرکز بوده و موضوع شناسایی گروه‌های گره‌های تأثیرگذار که به صورت جمعی بیشترین تأثیر را دارند، به عنوان چالشی مهم برای تحقیقات آینده باقی می‌ماند. شناسایی چنین گروه‌هایی می‌تواند در کاربردهایی مانند کنترل اپیدمی‌ها یا گسترش اطلاعات بهینه، اهمیت ویژه‌ای داشته باشد.

برای توسعه بیشتر این چارچوب، استفاده از ترکیب شاخص‌های مرکزی بیشتر و بهره‌گیری از شبکه‌های عصبی گراف به منظور بهبود نمایش گره‌ها پیشنهاد می‌شود. این رویکردها می‌توانند به درک بهتر تعاملات بین گره‌ها و شناسایی دقیق‌تر گره‌های کلیدی کمک کنند. همچنین، تحقیقات آینده می‌تواند به ایجاد روش‌هایی برای کاهش بیشتر زمان محاسباتی و افزایش مقیاس‌پذیری مدل بپردازد. یکی از چالش‌های اصلی این پژوهش نیز، محدودیت در منابع تولید برچسب و زمان اجرای شبیه‌سازی‌ها بود. برای تولید برچسب‌ها، شبیه‌سازی مدل SIR برای هر گره  $1000$  بار تکرار شد. این مقدار ثابت برای تمام گراف‌ها در نظر گرفته شد، در حالی که تعداد دفعات شبیه‌سازی لازم می‌تواند برای هر گراف متفاوت باشد. برای برخی گراف‌ها این مقدار ممکن است ناکافی باشد و برای برخی دیگر بیش از حد لازم باشد. به همین دلیل، بهینه‌سازی تعداد دفعات شبیه‌سازی بر اساس نیازهای خاص هر گراف به صورت داینامیک می‌تواند به دقت بیشتر برچسب‌ها کمک کند.

یکی از رویکردهای پیشنهادی در این راستا، بررسی رفتار  $10$  درصد برتر گره‌ها در شبیه‌سازی‌ها است. این گره‌ها از اهمیت بیشتری برخوردارند، زیرا هدف اصلی ما شناسایی گره‌های با بالاترین تأثیرگذاری است. تحلیل دقیق‌تر این گره‌ها می‌تواند به تنظیم پویا و دقیق تعداد شبیه‌سازی‌های لازم کمک کردد و از مصرف غیرضروری منابع جلوگیری کند. این تغییر می‌تواند بهبود قابل توجهی در عملکرد مدل و دقت برچسب‌های تولید شده ایجاد کند.

## كتاب نامه

- [1] G. Zhao, P. Jia, A. Zhou, and B. Zhang, “Infgcn: Identifying influential nodes in complex networks with graph convolutional networks,” *Neurocomputing*, vol.414, pp.18–26, 2020.
- [2] Y. Ou, Q. Guo, J.-L. Xing, and J.-G. Liu, “Identification of spreading influence nodes via multi-level structural attributes based on the graph convolutional network,” *Expert Systems with Applications*, vol.203, p.117515, 2022.
- [3] E. Yu, D. Chen, Y. Fu, and Y. Xu, “Identifying critical nodes in complex networks by graph representation learning,” *CoRR*, vol.abs/2201.07988, 2022.
- [4] C. Liu, T. Cao, and L. Zhou, “Learning to rank complex network node based on the self-supervised graph convolution model,” *Knowledge-Based Systems*, vol.251, p.109220, 2022.
- [5] W. Ahmad, B. Wang, and S. Chen, “Learning to rank influential nodes in complex networks via convolutional neural networks,” *Applied Intelligence*, vol.54, pp.1–19, 03 2024.
- [6] A.-L. Barabasi and R. Albert, “Albert, r.: Emergence of scaling in random networks. science 286, 509-512,” *Science (New York, N.Y.)*, vol.286, pp.509–12, 11 1999.
- [7] C. Castellano and R. Pastor-Satorras, “Thresholds for epidemic spreading in networks,” *Physical Review Letters*, vol.105, Nov. 2010.

- [8] Y. Feld and A. K. Hartmann, “Large deviations of a susceptible-infected-recovered model around the epidemic threshold,” *Physical Review E*, vol.105, Mar. 2022.

## پیوست الف

در این پیوست، جزئیات مربوط به یکی از الگوریتم‌های اصلی استفاده شده در این پروژه و همچنین لینک به کدهای متن‌باز پروژه ارائه می‌شود. کدهای پروژه به صورت متن‌باز در مخزن گیتهاب زیر قابل دسترسی هستند:<sup>۱</sup>

### الگوریتم

یکی از الگوریتم‌های اساسی مورد استفاده در این پروژه، مدلی برای تحلیل تأثیرگذاری گره‌ها در شبکه‌های پیچیده است که با استفاده از شبیه‌سازی مدل SIR بر روی شبکه‌های گرافی، نتایج مرتبط با مقیاس‌های تأثیرگذاری و مقادیر مرتبط با گره‌های آلوده شده را محاسبه می‌کند.

مراحل کلی الگوریتم:

۱. دریافت گراف از ورودی

۲. محاسبه مقادیر  $\beta$  برای گراف

۳. اجرای مدل SIR برای هر گره با مقادیر مختلف  $\beta$

۴. ذخیره نتایج مربوط به تأثیرگذاری

در ادامه شبه کد <sup>۱</sup> این الگوریتم آورده شده است.

<sup>1</sup><https://github.com/neginkheirmand/influential-node-ranking-in-complex-networks>

---

### Algorithm 1 Generate SIR Labels for a Graph

---

```

1: Input: Graph  $G$ , Number of  $B$  values ( $num\_b$ ), Infected Nodes
2: Output: Affected scales and Infected scales for each node
3: function GET_B_VALUE(Graph  $G$ , Number of  $B$  values ( $num\_b$ ))
4:   Compute the degree for each node in  $G$ :  $degrees = [\text{deg for each node in } G]$ 
5:   Compute mean degree:  $mean\_degree = \text{mean}(degrees)$ 
6:   Compute mean of squared degrees:  $mean\_degree\_squared = \text{mean}(degrees^2)$ 
7:   Compute Epidemic Threshold:  $B\_Threshold = \frac{mean\_degree}{mean\_degree\_squared - mean\_degree}$ 
8:   Compute the range of  $B$  values:
9:      $B\_values = \text{linspace}(1 \times B\_Threshold, 2 \times B\_Threshold, num\_b)$ 
10:  Round the  $B$  values:  $B\_values = \text{round}(B\_values, 3)$ 
11:  return  $B\_values$ 
12: end function
13: function SIR(Graph  $G$ , Infected nodes,  $B\_values$ ,  $\gamma$ , Number of iterations, Number of steps)
14:   Initialize empty dictionaries:  $affected\_scales$ ,  $infected\_scales$ 
15:   for each  $B$  in  $B\_values$  do
16:     Initialize  $recovered\_sum = 0$ ,  $infected\_sum = 0$ 
17:     for each iteration in range( $num\_iterations$ ) do
18:       Initialize SIR model:  $model = \text{SIRModel}(G)$ 
19:       Set infection rate to  $B$ :  $model.set\_parameter('beta', B)$ 
20:       Set recovery probability to  $\gamma$ :  $model.set\_parameter('gamma', \gamma)$ 
21:       Set initial infected nodes:  $model.set\_initial\_status(infected)$ 
22:       for each step in range( $num\_steps$ ) do
23:         Execute one step of the model:  $model.iteration()$ 
24:         if All infected nodes are recovered or susceptible then
25:           Break the loop if no infected nodes are left
26:         end if
27:       end for
28:       Update  $recovered\_sum$  and  $infected\_sum$  with the final state
29:     end for
30:     Compute  $affected\_scale$  for current  $B$ :
31:      $affected\_scale = \frac{recovered\_sum}{num\_iterations \times num\_nodes}$ 
32:     Store the result:  $affected\_scales[B] = affected\_scale$ 
33:     Store infected nodes sum:  $infected\_scales[B] = infected\_sum$ 
34:   end for
35:   return  $affected\_scales$ ,  $infected\_scales$ 
36: end function
37: function SIR_OF_GRAPH(Graph path, Number of  $B$  values, Result path)
38:   Load the graph  $G$  from file:  $G = \text{read\_edgelist}(graph\_path)$ 
39:   Compute  $B\_values$  using get_B_Value( $G$ ,  $num\_b$ )
40:   for each node in  $G.nodes$  do
41:     Initialize infected set:  $infected = \{node : 1\}$ 
42:     Compute affected and infected scales using SIR( $G$ ,  $infected$ ,  $B\_values$ )
43:     Store the results in the paths
44:   end for
45:   return Final results
46: end function

```

---