# مقدمه ای بر داده های گراف


## چکیده

استخراج و مدیریت گراف ها به واسطه کاربرد های متعددی که در دامنه وسیعی از مسائل استخراج داده ها و اطلاعات در زیست شناسی محاسباتی، تحلیل داده های شیمی، کشف مواد دارویی و ایجاد شبکه های ارتباطی دارد، اخیرا به موضوعی مهم در حوزه تحقیق تبدیل شده است. الگوریتم های سنتی استخراج و مدیریت داده ها مانند خوشه ای[۱]، طبقه بندی[۲]، استخراج الگوهای متداول[۳] و شاخص گذاری[۴] اکنون به شکل گراف ها گسترش یافته اند. کتابی که در دست دارید شامل فصل هایی است که با دقت بسیار انتخاب شده اند به این منظور که موضوعات تحقیق بی شماری در حوزه مدیریت و استخراج داده ها به بحث گذاشته شود. علاوه بر این، کاربردهای مهم و قابل توجه استخراج داده های گراف در این کتاب گنجانده شده است. هدف این فصل ارائه شرح مختصری از انواع تکنیک های پردازش و استخراج گراف ها و ارائه گزارش از این مباحث است.



**واژه های کلیدی:** استخراج داده های گراف، مدیریت گراف


[۱] Clustering

[۲] Classification

[۳] Frequent Pattern Mining

[۴] Indexing

## ۱. مقدمه

در این فصل، مقدمه ای از مبحث مدیریت و استخراج داده ها و رابطه آن با سایر فصل های این کتاب ارائه خواهد شد. مسئله مدیریت گراف در حال حاضر کاربردهای فراوانی در دامنه وسیعی از حوزه های کاربردی مانند تحلیل داده های شیمی، زیست شناسی محاسباتی، شبکه های اجتماعی، تجزیه و تحلیل لینک های وب و شبکه های کامپیوتری پیدا کرده است. برنامه های کاربردی متنوع و گوناگون منجر به پیدایش گراف های متفاوتی شده است و چالش های متناظر با آن نیزکاملا متفاوتند. به عنوان مثال، گراف های داده های شیمی نسبتا ناچیزند اما برچسب های گره های مختلف (که از مجموعه کوچکی از فاکتور ها و عوامل استخراج شده اند) ممکن است بارها در یک مولکول (گراف) تکرار شود. به همین خاطر مسائلی پیش می آید که هم ریختی و هم شکلی گراف ها در برنامه های استخراج و مدیریت داده ها را درپی دارد. از سوی دیگر در حوزه های بسیار گسترده ای [۱۲،۲۱،۲۲] مثل شبکه جهانی (وب)، شبکه های کامپیوتری و شبکه های اجتماعی، برچسب گره ها (به عنوان مثال URLs) مجزاست، اما تعداد بسیار زیادی از آنها وجود دارد. اینگونه گراف ها دشوار و چالش انگیز نیز هستند زیرا توزیع مرتبه این گراف ها تا حد زیادی مخدوش است [۱۰] و این منجر به ایجاد دشواری هایی در طبقه بندی این گراف به صورت اختصاری می شود. در بسیاری موارد، گراف ها ممکن است دینامیک (پویا) و تکاملی باشند، به این معنا که ساختار گراف ممکن است در طول زمان به سرعت دچار تغییر و تحول شود. در اینگونه موارد، جنبه زمانی تحلیل شبکه بی نهایت جالب خواهد بود.

یکی از شاخه های بسیار نزدیک، داده های XML است. داده های پیچیده و نیمه سازماندهی شده اغلب به صورت اسناد XML ارائه می شود چرا که این نوع داده ها دارای قدرت القایی و گویایی ذاتی هستند. داده های XML معمولا به شکل گراف ارائه می شوند، به طوری که خاصیت ها همراه با مقادیرشان به صورت گره ها ارائه می شوند، و رابطه میان آنها به عنوان کران ها (لبه ها) تعریف می شوند. قدرت القایی و گویایی گراف ها و داده های XML در ازای هزینه ای به دست می آید، در حالیکه طراحی سیستم عامل هایی برای استخراج و مدیریت داده های سازماندهی شده بسیار دشوارتر است. طراحی الگوریتم های مدیریت و استخراج برای داده های XML به طراحی روش هایی موثر برای ارائه داده های گراف نیز کمک می کند زیرا این دو حوزه ارتباط نزدیکی با یکدیگر دارند.

این کتاب برای بررسی جنبه های گوناگون مدیریت و استخراج داده ها طراحی شده است و چکیده ای در اختیار سایر محققان این حوزه قرار می دهد. هسته اصلی این کتاب به سه بخش تقسیم می شود:

- **سازماندهی و مدیریت داده های گراف:** از آنجایی که گراف ها گونه ای پیچیده و القایی از داده ها را تشکیل می دهند، ما به روش هایی برای ارائه گراف ها در پایگاه های داده، بهره گیری از آنها و بررسی آنها نیاز داریم.

ما به بررسی مسئله طراحی زبان های استفهامی (پرسشگر) برای گراف ها پرداختیم [۱۴]، و نشان دادیم چگونه می توان به منظور بازیابی ساختارهایی خاص از گراف ها از این زبان ها استفاده کرد [۲۶]. همچنین طراحی ساختارهای بازیابی و شاخص گذاری برای داده های گراف را نیز بررسی کردیم. علاوه بر این، تعدادی از پرسش های تخصصی مانند تطبیق، جستجوی واژه های کلیدی و پرسش های قابلیت دسترسی [۴،۷،۲۴] در این کتاب مورد مطالعه و بررسی قرار گرفته است. در این کتاب خواهیم دید که طراحی شاخص برای نرم افزارهای اصلی و زیربنایی در مورد داده های سازماندهی شده بسیار حساس تر از داده های چند بعدی است. مسئله مدیریت داده های گراف به دامنه گسترده مدیریت داده های XML مربوط است. در صورت امکان، از حوزه داده های XML کمک خواهیم گرفت. نشان خواهیم داد که بعضی از این تکنیک ها ممکن است به منظور مدیریت گراف ها و طرح های گراف در حوزه های مختلف مورد استفاده قرار گیرند. همچنین بعضی از تکنیک هایی که اخیرا برای داده های گراف طراحی شده اند را نیز معرفی خواهیم کرد.

- **استخراج داده های گراف:** مانند سایر داده ها از جمله داده های سه بعدی یا داده های متنی، ما می توانیم مسائل استخراج داده های گراف را در نظر بگیریم و برای آن برنامه ریزی کنیم که این شامل تکنیک هایی مثل استخراج الگوهای متداول، خوشه ای و طبقه بندی است [۱،۱۱،۱۶،۱۸،۲۳،۲۵،۲۶،۲۸]. باید اشاره کنیم که این روش ها در حوزه گراف ها چالش های بیشتری به وجود می آورند زیرا ماهیت ساختاری داده ها قابلیت تفسیر و ارائه نتایج استخراج داده ها را با دشواری و مشکل مواجه می کند. البته این مشکل به دشواری ایجاد وضوح و گویایی بیشتر برای گراف ها مرتبط است.
- **کاربرد های گراف:** بسیاری از تکنیک هایی که در بالا مورد بحث و بررسی قرار گرفتند برای گراف های کلی و عام در شرایط و موارد ویژه است. با این وجود، حوزه های مرتبط با گراف ها بی نهایت متنوع و گوناگون است و این ممکن است به تفاوت های بی شماری در الگوریتم های طراحی شده برای چنین مواردی منتهی شود. به عنوان مثال، الگوریتم هایی که برای وب یا شبکه های اجتماعی طراحی شده اند باید با اندازه ای بسیار بزرگ ولی برچسب های مجزای گره ها برای گراف ها بازسازی شوند. از سوی دیگر، در الگوریتم هایی که برای داده های شیمی طراحی شده اند می بایست تکرار در برچسب گره ها در نظر گرفته شود، به همین شکل، بسیاری از گراف ها و منحنی ها ممکن است به اطلاعات تکمیلی و اضافی همراه با گره ها و کران ها نیاز داشته باشند. این تنوع و گوناگونی باعث می شود که نرم افزارهای مختلف با چالش بیشتری رو به رو شوند. به علاوه تکنیک های کلی و عمومی که پیشتر مورد بحث قرار گرفت ممکن است برای حوزه های مختلف نرم افزاری کاربردهای متفاوتی داشته باشند. از این رو، برای اینکه به این موارد و گونه های متفاوت بپردازیم فصل های مختلفی را درکتاب گنجاندیم. در فصل های این کتاب به بررسی برنامه های کاربردی

مرتبط با وب، شبکه های اجتماعی، مکان یابی عیب یاب های نرم افزار، داده های شیمی و زیست شناسی خواهیم پرداخت.

یکی از اهداف این کتاب، ارائه نمای کلی فراگیری از موارد و ابزار در حوزه مدیریت و استخراج داده ها است. کتاب حاضر در ابتدا به معرفی چند فصل می پردازد، سپس انواع مختلفی از الگوریتم های استخراج گراف را با جزییات و به طور مفصل مورد بحث و بررسی قرار می دهد.

## ۲. برنامه های کاربردی مدیریت و استخراج داده ها

در این بخش، سازمان بندی بخش های مختلف کتاب مورد بحث قرار خواهیم داد. برنامه های مختلف را بررسی خواهیم کرد و فصل هایی که به بحث در مورد این برنامه ها می پردازد را معرفی خواهیم کرد. در دو بخش اول، مقدمه ای بر حوزه استخراج گراف ها به عنوان یک ارزیابی کلی خواهیم آورد. این بخش (بخش اول) مقدمه ای مختصر در مورد حوزه مدیریت داده ها و سازمان بندی این کتاب ارائه می دهد. فصل دو یک ارزیابی و بازبینی کلی است که مسائل کلیدی و الگوریتم های هر یک از حوزه ها را مورد بحث قرار می دهد. هدف دو فصل اول این است که رئوس کلی این حوزه را بدون وارد شدن به جزییات در اختیار خواننده قرار دهد. فصول بعدی به حوزه های مختلف استخراج داده ها خواهند پرداخت. در ادامه به این موارد و موضوعات خواهیم پرداخت.

### ویژگی های طبیعی گراف های واقعی.

به منظور درک تکنیک های مختلف مدیریت و استخراج که در این کتاب ارائه شده، لازم است بدانیم که گراف های واقعی در عمل چگونه اند. گراف هایی که در برنامه هایی با مقیاس بزرگ مانند شبکه وب و شبکه های اجتماعی به کار گرفته می شوند دارای ویژگی های زیادی مثل توزیع قدرت و توان [۱۰]، سادگی، و قطر های کوچک هستند [۱۹]. این ویژگی ها نقشی کلیدی در طراحی الگوریتم های کارامد مدیریت و استخراج برای گراف ها بازی می کنند. از این رو، در بخش ابتدایی کتاب به بحث درباره این ویژگی ها خواهیم پرداخت. به علاوه، تکامل گراف های دینامیک مانند شبکه های اجتماعی نشان دهنده ویژگی های جالب توجهی است. از جمله: تراکم[۵]، و قطرهای کوچک شونده[۶] [۱۹]. علاوه بر این، از آنجایی که مطالعه و بررسی الگوریتم استخراج داده ها نیازمند طراحی مجموعه ای از مولد های کارامد و موثر گراف است، بررسی روش هایی برای ایجاد مولد های مناسب می تواند مفید واقع شود[۳]. بی تردید آگاهی و شناختی که با مطالعه ویژگی های گراف ها در حوزه های

[۵] Densification

[۶] Shrinking Diameters

واقعی  به دست آوردیم می تواند در طراحی مدل هایی برای ایجاد مولدهای کارامد و مفید تاثیر بسزایی داشته باشد. فصل ۳، قواعد و قوانین گراف های حقیقی شبکه های عظیم و تعدادی از تکنیک های تولید ساختگی و ترکیبی گراف ها را مورد بحث قرار می دهد.

### زبان های استفهامی[۷] و شاخص گذاری برای گراف ها.

به منظور استفاده موثر از نرم افزارهای مدیریت گراف به زبان های استفهامی که مکان وضوح و گویایی برای مدیریت و استفاده از داده های ساختاری را فراهم می اورند، نیاز داریم. به علاوه، چنین زبان هایی باید به نحو موثر و کارامدی به اجرا درایند. در فصل ۴، انواعی از زبان های استفهامی برای گراف ها ارائه شده است.

مسئله دیگر به دست یابی موثر به اطلاعات مخفی به منظور پاسخ دادن به پرسش ها و تردیدها مربوط است. به همین خاطر، بررسی و مطالعه طراحی ساختارهای شاخص گذاری برای گراف ها مفید خواهد بود. تکنیک های کلی برای شاخص گذاری موثر گراف ها در فصل ۵ ارائه شده است. در حالی که فصل ۵ منحصرا بر حوزه گراف تمرکز کرده است، اشاره کردیم که بسیاری از تکنیک های شاخص گذاری برای حوزه XML می تواند در حوزه گراف ها نیز موثر باشد. فصل دو بعضی از پیوستگی ها و ارتباط ها میان شاخص گذاری XML و شاخص گذاری گراف ها را مورد بررسی قرار می دهد. علاوه بر مسائلی مانند جستجوی شباهت ها[۸]، که معمولا برای مجموعه داده های چند گراف طراحی شده است، ساختارهای گراف با طراحی تعدادی از پرسش های مربوط به یک گراف بزرگ مجزا متناسب و سازگارند. در این گونه موارد، ممکن است یک گراف مجزا داشته باشیم، اما خواستار تعیین ویژگی های درون-گره در گراف خواهیم بود. این گونه مسائل و پرسش ها اغلب اوقات در چهارچوب شبکه های اجتماعی و وب پدید می آید. نمونه هایی از این پرسش ها شامل قابلیت دسترسی و پرسش های بعد-محور است. [۲،۴،۷،۲۴]. این گونه پرسش ها در مبنای رفتار ابعاد درون گره ای در ساختار شبکه پدید آمده، و اغلب دشوار و چالشی به نظر می رسند زیرا گراف اصلی ممکن است روی دیسکت ذخیره شده باشد. در فصل ۶، آثار و نوشته های مربوط به پردازش مسائل و پرسش های قابلیت دسترسی بررسی خواهد شد.

### تطبیق گراف[۹].

تطبیق گراف ها مسئله ای جدی است که در زمینه بسیاری از انواع نرم افزارها مانند تطبیق، نصب گراف و سایر نرم افزارهای تجاری به وجود امده است [۹]. در مسئله تطبیق گراف، ما دو گراف داریم و سعی می کنیم طراحی

---

[۷] Query Languages

[۸] Similarity Search

[۹] Graph Matching

از گروه های بین دو گراف تعیین کنیم به صورتی که کران و/یا انطباق برچسب حفظ شود. تطبیق گراف به طور سنتی در آثار نظری در مورد مسئله هم ریختی گراف ها مورد تحقیق و مطالعه قرار گرفته است. با این وجود، در زمینه برنامه های کاربردی عملی، ممکن است تطبیق دقیق و درست دو گراف امکان پذیر نباشد. علاوه بر این، بسیاری از گونه های عملی و عینی مسئله امکان وجود اطلاعات و دانش محدود و ناقص در مورد تطبیق گروه های مختلف را در نظر می گیرد. از این رو، به تحقیق درباره تکنیک های دقیق و غیر دقیق تطبیق گراف ها می پردازد.

**جستجوی کلید واژه در گراف ها.**

در مسئله جستجوی کلید واژه ها خواستار تعیین گروه های کوچکی از گره های دارای لینک های پیوسته هستیم که به کلید واژه های ویژه ای مرتبط باشد [۱۵]. به عنوان مثال، یک گراف وب یا یک شبکه اجتماعی ممکن است به عنوان یک گراف بزرگ در نظر گرفته شود [۲۱،۲۲]، که هر گره آن ممکن است شامل مقادیر زیادی داده های متنی باشد. به رغم اینکه جستجوی کلید واژه ها با توجه به متن درون گره ها تعیین می شود، باید به این نکته اشاره کنیم که ساختار اتصال نیز نقشی کلیدی در تعیین مجموعه مناسبی از گره ها ایفا می کند. اطلاعات موجود در متن ها و ساختار و ترکیب اتصال یکدیگر باز تایید می کنند و این عمل منجر به نتایجی با کیفیت بالاتر می شود. جستجوی کلید واژه ها سطح مشترک ساده اما کاربر پسند برای بازیابی اطلاعات روی شبکه ایجاد می کند، همچنین، به اثبات رسیده است که جستجوی کلید واژه ها روشی موثر و کارامد برای جستجوی داده های موجود در ساختارهای پیچیده است. از آنجایی که بسیاری از مجموعه داده های موجود در زندگی واقعی به صورت جدول، درخت و گراف سازمان بندی شده اند، جستجوی کلید واژه ها برای اینگونه داده ها اهمیت ویژه و فزاینده ای یافته است، و توجه زیادی چه در زمینه پایگاه داده ها و چه جوامع IR به خود معطوف کرده است. لازم است تکنیک هایی برای جستجوی کلید واژه ها طراحی شوند که معانی پرسش ها، دقت رتبه بندی و کارایی و بازدهی پرسش ها را تقویت کنیم. فصل ۸ یک ارزیابی فراگیر و همه جانبه از تکنیکهای جستجوی کلید واژه ها در گراف ها است.

**خوشه بندی گراف و انتخاب زیر منحنی متراکم.**

مسئله خوشه بندی به دو صورت مختلف ظاهر میشود:

- در وهله اول، می خواهیم خوشه های گره متراکم در یک گراف بزرگ را تعیین کنیم. این مسئله در صورت وجود تعدادی از نرم افزاری ها مانند تقسیم بندی گراف[۱۰] و برش مینیمم[۱۱] پدید می آید. تعیین نواحی متراکم در گراف از دیدگاه تعداد نرم افزار های مختلف در شبکه های اجتماعی، خوشه بندی گراف وب و جمع بندی می تواند تبدیل به مسئله ای جدی و بغرنج شود. به ویژه، اکثر فرم های جمع بندی گراف نیازمند تعیین نواحی متراکم در گراف های اصلی است. تعدادی از تکنیک ها [۱۱،۱۲،۲۳] برای خوشه بندی گراف های متراکم در آثار و نوشته ها طراحی شده اند.
- در مرحله بعد گراف های چندگانه ای داریم که هر یک از آنها احتمالا دارای اندازه ای متوسط هستند. در این حالت، می خواهیم گراف ها را خوشه بندی کنیم. فاصله بین گراف ها براساس تابع تشابه ساختاری مثلا مسافت ویرایش و تصحیح تعیین می شود. به عنوان راهی دیگر ممکن است آن را براساس سایر ویژگی های کلی مانند عضویت الگوهای پر تکرار در گراف تعیین کرد. چنین تکنیک هایی به ویژه برای گراف های حوزه XML که ذاتا به صورت موضوعی بیان می شوند، سودمند خواهد بود. یکی از این روش های خوشه بندی داده های XML در [۱] مورد بحث قرار گرفته است.

در فصل ۹، هر دو روشی که در بخش قبل برای خوشه بندی گراف ها ذکر شد مورد بررسی قرار گرفته است. یکی از مسائل کاملا مرتب با خوشه بندی گراف ها، انتخاب یا استخراج زیر نمونارهای متراکم است. در حالی که مسئله خوشه بندی به طور سنتی به عنوان تقسیم بندی دقیق گره ها تعریف می شود. مسئله استخراج زیر گراف های متراکم گونه ای از این مسئله است که ممکن است با زیرگراف های مترکم همپوشانی کند. علاوه بر این، ممکن است بسیاری از گره ها در هیچ یک از مولفه های متراکم قرار نگیرد. مسئله زیرگراف های متراکم اغلب در چارچوب استخراج الگوهای پرتکرار از مجموعه داده های چند گراف مورد مطالعه قرار می گیرد. سایر گونه ها و موضوعات شامل مسئله حضور تکراری زیرگراف ها در یک گراف مجزا یا در گراف های چندگانه ای است. این مسائل در فصل ۱۰ مورد بحث قرار گرفته است. موضوعاتی که در فصل های ۹ و ۱۰ مورد بحث قرار می گیرند کاملا مرتبط هستند و نمای کلی این حوزه را در اختیار خواننده قرار می دهد.

### طبقه بندی گراف.

در صورت خوشه بندی گراف، مسئله طبقه بندی گراف به دو حالت ظاهر می شود. حالت مربوط به طبقه بندی راس است که سعی می کنیم گره های یک گراف مجزا را براساس داده های آموزشی برچسب گذاری کنیم: این

[۱۰] Graph-Partitioning

[۱۱] Minimum Cut

گونه مسائل بر اساس تعیین ویژگی های مطلوب گره ها با استفاده از داده های آموزشی استوار است. نمونه هایی از این روش ها و متدها را می توانید در [۱۶،۱۸] پیدا کنید. حالت دوم زمانی روی می دهد که سعی می کنیم تمام گراف ها را به صورت موضوعی برچسب بزنیم. حالت دوم در صورتی روی می دهد که گراف های بزرگی مانند شبکه های اجتماعی وجود داشته باشند. در حالی که حالت دوم در بسیاری از شرایط مانند طبقه بندی ترکیب های شیمیایی و زیستی یا داده های XML [۲۸] روی می دهد. فصل ۱۱ تعدادی از الگوریتم های مختلف برای طبقه بندی گراف ها را مورد بحث قرار می دهد.

### استخراج الگوهای پرتکرار از گراف ها.

مسئله استخراج الگوهای پر تکرار در گراف ها دشوارتر از داده های اجرایی استاندارد است زیرا تمام الگوهای پر تکرار گراف ها به یک اندازه مرتبط و مناسب نیستند. به ویژه، الگوهایی که پیوستگی و ارتباط زیادی دارند مناسب ترند. درصورت وجود داده های اجرایی ممکن است چندین مقیاس مختلف تعریف شود به این منظور که تعیین شود کدام گراف ها معنی دار ترند. در صورت وجود گراف ها، محدودیت های ساختاری می تواند مسئله را جالب تر کند و در صورت وجود داده های اجرایی، بسیاری از حالت های استخراج الگوهای گراف مانند تعیین الگوهای نزدیک[۱۲] یا الگوهای معنی دار[۱۳] [۲۵،۲۶] انواع مختلفی از چشم اندازهای این حوزه را ارائه می دهد.

مسئله استخراج الگوهای پرتکرار اهمیت ویژه ای برای حوزه گراف دارد زیرا نتایج نهایی الگوریتم ها یک نمای کلی از ساختارهای مهم در مجموعه داده های اصلی فراهم می آورد که ممکن است در سایر برنامه های کاربردی از قبیل شاخص گذاری [۲۷] قابل استفاده باشند. فصل ۱۲ یک ارزیابی فراگیر از الگوریتم های مختلف برای استخراج الگوهای پرتکرار در گراف ها ارائه می دهد.

### گروه بندی الگوریتم ها[۱۴] برای گراف ها.

بسیاری از کاربردهای گراف از جمله آنهایی که در مخابرات و شبکه های اجتماعی کاربرد دارند گروه های پیوسته ای از کران ها را ایجاد می کنند. این گونه برنامه ها چالش های خاص خود را دارند زیرا نمی توان کل گراف را در یک حافظه اصلی یا روی دیسکت ذخیره و نگه داری کرد. این موضوع باعث ایجاد محدودیت های بسیار زیادی برای الگوریتم های اصلی می شود، هنگامی که الزام تک گذرگاه بودن برای الگوریتم های گروه بندی شده در این

[۱۲] Closed Patterns

[۱۳] Significant Patterns

[۱۴] Streaming Algorithms

مورد اعمال می شود. علاوه بر این، بررسی ویژگی های ساختاری گراف های اصلی بسیار دشوار است زیرا به سختی می توان دیدگاه جهانی در مورد گراف را در موارد گروه بندی شده ، طرح ریزی کرد. فصل ۱۳ تعدادی از برنامه های گروه بندی شده برای چنین گروه هایی از کران ها را مورد بررسی قرار می دهد. این فصل به این بحث می پردازد که چگونه گروه های گراف را می توان به شیوه ای مختص برنامه های کاربردی جمع بندی کرد، به طوری که ویژگی های ساختاری مهم گراف قابل بررسی باشد.

**استخراج داده های شخصی–محافظت شده.**

در بسیاری از نرم افزار ها از جمله شبکه های اجتماعی، حفظ محدوده حفاظت شده گره ها در شبکه های اجتماعی اهمیت زیادی دارد. هویت زدایی سالم از گره ها در طی عرضه یک ساختار شبکه به تنهایی کافی نیست، زیرا ممکن است رقیبی از اطلاعات پیش زمینه درباره گره های شناخته شده استفاده کند. به این منظور که سایر گره ها را دوباره شناسایی کند [۱۷]. محدوده حفاظت شده گراف به ویژه از آن جهت چالش انگیز است که می توان اطلاعات پیش زمینه درباره بسیاری از ویزگی های ساختاری از قبیل رتبه های گره یا فواصل ساختاری را برای بالا بردن حمله های-شناسایی گره ها مورد استفاده قرار داد [۱۷،۱۳]. اخیرا تعدادی از تکنیک ها در آثار مکتوب پیشنهاد شده اند که از افزایش، حذف یا تعویض گره ها استفاده می کند به این منظور که اینگونه ویزگی های ساختاری در جهت اهداف حفظ حریم و حدود-محافظت پنهان شوند [۲۰،۲۹]. نکته کلیدی این است که ویژگی های ساختاری تعیین هویت را بدون از دست دادن فعالیت ساختاری گراف پنهان کنیم. فصل ۱۴ به بررسی چالش های حدود گراف و انواع الگوریتم های قابل استفاده در پردازش حفاظت شده گراف ها می پردازد.

**برنامه های کاربردی شبکه.**

از آنجایی که وب معمولا به شکل یک گراف سازمان بندی می شود، بسیاری از این گراف ها نیازمند الگوریتم های استخراج و مدیریت گراف هستند. یکی از نمونه های کلاسیک در این مورد شبکه های اجتماعی است که ساختار اتصال به شکل گراف تعریف شود. برنامه های متداول شبکه سازی اجتماعی نیازمند تعیین نواحی مهم و جالب در گراف ها مانند مناطق متراکم هستند. ابزرسی منطقه یکی از برنامه های کاربردی مستقیم در مسئله خوشه بندی است، در حالی که مستلزم تعیین نواحی متراکم گراف های اصلی است. اکثر برنامه های دیگر مانند تحلیل و تفسیر بلاگ ، تحلیل گراف شبکه و تحلیل رتبه صفحه نیازمند استفاده از الگوریتم های استخراج گراف است. فصل ۱۵ یک چشم انداز کلی و همه جانبه از تکنیک های استخراج گراف برای برنامه های شبکه ارائه می دهد. از آنجایی که ایجاد شبکه های اجتماعی حوزه ای بسیار مهم است که به سادگی نمی توانم آن را در چارچوب

فصلی مجزا در برنامه های شبکه گنجاند، فصلی ویژه به ساخت شبکه های اجتماعی اختصاص دادیم. برنامه های استخراج گراف برای ساخت شبکه های اجتماعی در فصل ۱۶ مورد بحث قرار می گیرد.

### مکان یابی عیب یاب های نرم افزار[۱۵].

برنامه های نرم افزاری را می توان به صورت گراف ها ارائه کرد به طوری که گردش ابزارهای کنترل پدید می آیند. اینگونه دستکاری ها را می توان در قالب ساختارهای گراف که معرف این گردش ابزارهای کنترل است، توجیه کرد. از این رو مکان یابی عیب یاب های نرم افزار یک برنامه ساده در الگوریتم های استخراج گراف است که ساختار گراف گردش ابزارهای کنترل را به منظور یقین و ایزوله کردن عیب ها در برنامه اصلی مورد مطالعه قرار می دهد، فصل ۱۸ یک ارزیابی و بازبینی همه جانبه از تکنیک های مکان یابی عیب یاب های نرم افزار ارائه می دهد.

### داده های شیمیایی و زیست شناسی.

ترکیب شیمیایی را می توان به شکل ساختارهای گراف تعریف کرد به طوری که اتم ها نشان دهنده گره ها و پیوندها نشان دهنده حلقه های اتصال باشند. در صورت تمایل می توان از سطح بالاتری برای بیان گراف ها استفاده کرد به طوری که واحدهای مولکولی معرف گره ها، پیوند های بین آنها بیانگر اتصالات باشد. به عنوان مثال در مورد داده های زیست شناسی آمینو اسیدها نشان دهنده گره ها، پیوند های بین آمینو اسیدها بیانگر اتصال ها می باشند. داده های شیمیایی و زیست شناسی به طور ذاتی متفاوتند، از این نظر که گراف های متناظر با داده های زیست شناسی بسیار بزرگ و مستلزم تکنیک های مختلفی هستند که بیشتر برای کارهای بزرگ مناسب هستند. بنابراین، دو فصل جداگانه را به موضوع اختصاص دادیم. در فصل ۱۸، روش های استخراج ترکیب های زیستی ارائه شده است. تکنیک های استخراج ترکیب های شیمیایی در فصل ۱۹ ارائه شده است.

## ۳. خلاصه

این کتاب مقدمه ای بر مسئله مدیریت و استخراج داده های گراف ارائه می دهد. ما تکنیک های کلیدی برای مدیریت و استخراج مجموعه داده های گراف ارائه می دهیم و نشان خواهیم داد که این تکنیک ها در دامنه وسیعی از برنامه های کاربردی مانند وب ، شبکه های اجتماعی ، داده های زیست شناسی ، داده های شیمیایی و مکان یابی عیب یاب های نرم افزار سودمند خواهد بود. کتاب همچنین بعضی از آخرین روش ها برای استخراج گراف

[۱۵] Software Bug Localization

های انبوه و بزرگ و قابلیت استفاده از آنها در حوزه های مختلف را بیان می کند. بعضی از روش ها در استخراج داده ها، حوزه های سازنده ای برای تحقیق در مورد برنامه های کاربردی آینده هستند:

- قابلیت رتبه بندی یکی از نیازهای جدید در برنامه های استخراج گراف است برنامه هایی از قبیل وب و شبکه های اجتماعی در مورد گراف های بزرگی تعریف می شوند که امکان ذخیره کران های اصلی آن روی حافظه یا حتی دیسکت وجود ندارد. در حالی که الگوریتم های گراف-نظری به طور مشروح در آثار مربوطه مورد مطالعه قرار گرفته اند. این تکنیک ها فرض را بر این می گذارند که گراف ها را می توان در حافظه اصلی نگهداری کرد و به همین خاطر در مورد اطلاعات روی دیسکت مفید نخواهد بود. به این دلیل که مراجعه به دیسکت ممکن است منجر به دستیابی تصادفی به کران های اصلی شود که در عمل ناکارامد خواهد بود. این قضیه به فقدان قابلیت رتبه بندی الگوریتم های اصلی نیز منتهی می شود.
- اکثر برنامه های ساخت شبکه های ارتباطی و اجتماعی مجموعه های بزرگی از کران ها را ایجاد می کنند که به طور مداوم و با بازگشت زمان وارد می شوند. اینگونه نرم افزارهای دینامیک مستلزم پاسخ سریع پرسش ها در مورد تعدادی از برنامه های سنتی مانند مسئله کوتاهترین مسیر و پرسش های مربوط به قابلیت اتصال است. اینگونه پرسش ها چالشی بزرگ است، زیرا پیش-ذخیره حجم زیادی از داده ها برای تجزیه و تحلیل آینده امکان پذیر نیست. بنابراین باید تکنیک های موثری برای فشرده کردن ساختارهای گراف برای تحلیل های آینده طراحی شود.
- تعدادی از برنامه ها و پیشرفت های اخیر در استخراج داده ها مانند استخراج داده های محافظت کننده و داده های نا مشخص باید در چارچوب حوزه گراف ها مورد مطالعه و تحقیق قرار بگیرد. به عنوان مثال، شبکه های اجتماعی به شکل گراف ها سازمانبندی می شوند و برنامه های محافظت به ویژه در این چارچوب خاص اهمیت پیدا می کنند. این گونه نرم افزار ها بسیار چالش انگیز هستند به این خاطر که برای حوزه وسیعی از گره ها تعریف شده اند.

این کتاب به مطالعه تعدادی از مسائل مهم در حوزه گراف در قالب برنامه های گراف و ساخت شبکه می پردازد. همچنین ، در این کتاب بعضی از گرایش های اخیر برای برنامه های استخراج گراف را معرفی می کنیم.

# مراجع


[1] C. Aggarwal, N. Ta, J. Feng, J. Wang, M. J. Zaki. XProj: A Framework for Projected Structural Clustering of XML Documents, KDD Conference,2007.

[2] R. Agrawal, A. Borgida, H.V. Jagadish. Efficient Maintenance of transitive relationships in large data and knowledge bases, ACM SIGMOD Conference, 1989.

[3] D. Chakrabarti, Y. Zhan, C. Faloutsos R-MAT: A Recursive Model for Graph Mining. SDM Conference, 2004.

[4] J. Cheng, J. Xu Yu, X. Lin, H. Wang, and P. S. Yu, Fast Computing Reachability Labelings for Large Graphs with High Compression Rate, EDBT Conference, 2008.

[5] J. Cheng, J. Xu Yu, X. Lin, H. Wang, and P. S. Yu, Fast Computation of Reachability Labelings in Large Graphs, EDBT Conference, 2006.

[6] E. Cohen. Size-estimation framework with applications to transitive closure and reachability, Journal of Computer and System Sciences, v.55 n.3, p.441-453, Dec. 1997.

[7] E. Cohen, E. Halperin, H. Kaplan, and U. Zwick, Reachability and distance queries via 2-hop labels, ACM Symposium on Discrete Algorithms, 2002.

[8] D. Cook, L. Holder, Mining Graph Data, John Wiley & Sons Inc, 2007.

[9] D. Conte, P. Foggia, C. Sansone, and M. Vento. Thirty years of graph matching in pattern recognition. Int. Journal of Pattern Recognition and Artificial Intelligence, 18(3):265–298, 2004.

[10] M. Faloutsos, P. Faloutsos, C. Faloutsos, On Power Law Relationships of the Internet Topology. SIGCOMM Conference, 1999.

[11] G. Flake, R. Tarjan, M. Tsioutsiouliklis . Graph Clustering and Minimum Cut Trees, Internet Mathematics, 1(4), 385–408, 2003.

[12] D. Gibson, R. Kumar, A. Tomkins, Discovering Large Dense Subgraphs in Massive Graphs, VLDB Conference, 2005.

[13] M. Hay, G. Miklau, D. Jensen, D. Towsley, P. Weis. Resisting Structural Re-identification in Social Networks, VLDB Conference, 2008.

[14] H. He, A. K. Singh. Graphs-at-a-time: Query Language and Access Methods for Graph Databases. InProc. Of SIGMOD ’08, pages 405–418, Vancouver, Canada, 2008.

[15] H. He, H. Wang, J. Yang, P. S. Yu. BLINKS: Ranked keyword searches on graphs. InSIGMOD, 2007.

[16] H. Kashima, K. Tsuda, A. Inokuchi. Marginalized Kernels between Labeled Graphs, ICML, 2003.

[17] L. Backstrom, C. Dwork, J. Kleinberg. Wherefore Art Thou R3579X? Anonymized Social Networks, Hidden Patterns, and Structural Steganography. WWW Conference, 2007.

[18] T. Kudo, E. Maeda, Y. Matsumoto. An Application of Boosting to Graph Classification, NIPS Conf.2004.

[19] J. Leskovec, J. Kleinberg, C. Faloutsos. Graph Evolution: Densification and Shrinking Diameters. ACM Transactions on Knowledge Discovery from Data (ACM TKDD), 1(1), 2007.

[20] K. Liu and E. Terzi. Towards identity anonymization on graphs. ACM SIGMOD Conference 2008.

[21] R. Kumar, P Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal. The Web as a Graph. ACM PODS Conference, 2000.

[22] S. Raghavan, H. Garcia-Molina. Representing web graphs. ICDE Conference, pages 405-416, 2003.

[23] M. Rattigan, M. Maier, D. Jensen: Graph Clustering with Network Sructure Indices. ICML, 2007.

[24] H. Wang, H. He, J. Yang, J. Xu-Yu, P. Yu. Dual Labeling: Answering Graph Reachability Queries in Constant Time. ICDE Conference, 2006.

[25] X. Yan, J. Han. CloseGraph: Mining Closed Frequent Graph Patterns, ACM KDD Conference, 2003.

[26] X. Yan, H. Cheng, J. Han, and P. S. Yu, Mining Significant Graph Patterns by Scalable Leap Search, SIGMOD Conference, 2008.

[27] X. Yan, P. S. Yu, and J. Han, Graph Indexing: A Frequent Structure-based Approach, SIGMOD Conference, 2004.

[28] M. J. Zaki, C. C. Aggarwal. XRules: An Effective Structural Classifier for XML Data, KDD Conference, 2003.

[29] B. Zhou, J. Pei. Preserving Privacy in Social Networks Against Neighborhood Attacks. ICDE Conference, pp. 506-515, 2008.