These algorithms were modified and extended to enable additional options: drawing with constraints (optimizing the selected part of the network, fixing some vertices to predefined positions, using values of edges as similarities or dissimilarities), drawing in 3D space. Pajek also provides tools for manual editing of graph layouts. The values of vectors can be used to determine several elements of network display such as X, Y, Z coordinates and the size of the vertex shape. The partition can be represented graphically by the color and shape of vertices. The values of edges can also be represented by thickness and/or color. Pajek also supports drawing sequences of networks in its Draw window, and exports sequences of networks in suitable formats that can be examined with special 2D or 3D viewers (such as SVG and Mage). Pictures in SVG can be further controlled using support written in Javascript.

Interfaces

Pajek also supports some non-native input formats: UCINET DL files; chemical MDLMOL and BS; and genealogical GEDCOM. The layouts can be exported in the following output graphic formats that can be examined by special 2D and 3D viewers: Encapsulated PostScript (EPS), Scalable Vector Graphics (SVG), VRML, MDLMOL/chime, and Kinemages (Mage). The main window menu Tools enables export of Pajek's data to statistical programs R and SPSS. In the Tools menu, the user can prepare calls to her/his favorite viewers and other tools. It is also possible to run Pajek (+macros) from other programs (R, Ucinet, and others).

This presentation of Pajek is a shortened and updated version of the chapter V. Batagelj, A. Mrvar. *Pajek—Analysis and Visualization of Large Networks,* in Jünger, M., Mutzel, P. (Eds.) *Graph Drawing Software,* pp 77-103. Springer, Berlin, 2003
This work was partially supported by the Ministry of Education, Science and Sport of Slovenia, Projects J1-8532 and Z5-3350.

Vladimir Batagelj / Andrej Mrvar

Pajek – Ein Programm zur Analyse und Visualisierung großer Netzwerke

Pajek ist ein unter Windows laufendes Programm zur Analyse und Visualisierung von großen Netzwerken mit Tausenden, ja, Millionen von Knoten (Vertices). „Pajek" ist das slowenische Wort für „Spinne". Die neueste Version von Pajek ist für nicht-kommerzielle Zwecke frei unter *http://vlado.fmf.uni-lj.si/pub/networks/pajek/* erhältlich.
Die Entwicklung von Pajek begann im November 1996. Das Programm ist in Delphi (Pascal) geschrieben. Einige Prozeduren hat Matjaz Zaversnik beigetragen. Hauptmotivation für die Entwicklung von Pajek war die Beobachtung, dass zahlreiche Quellen großer Netzwerke bereits in maschinenlesbarer Form vorliegen. Pajek sollte Werkzeuge zur Analyse und Visualisierung von solchen Netzwerken zur Verfügung stellen: von Kooperationsnetzwerken, organischen Molekülen in der Chemie, Netzwerken von Protein-Rezeptor-Wechselwirkungen, Genealogien, Internet-Netzwerken, Zitiernetzwerken, Diffusionsnetzwerken (AIDS, Nachrichten,

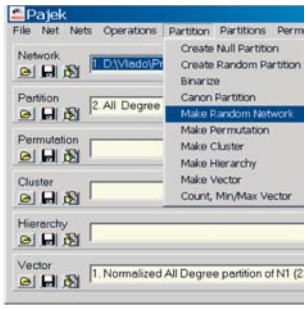
Innovationen), Datenmining (Two-Mode-Netzwerken) usw. Siehe dazu auch die Sammlung großer Netzwerke unter: *http://vlado.fmf.uni-lj.si/pub/networks/data/*
Das Design von Pajek beruht auf unseren früheren Erfahrungen mit der Entwicklung von Graphen-Strukturen und den Algorithmenbibliotheken Graph und X-Graph, einer Reihe von Netzwerkanalyse und -visualisierungsprogrammen, STRAN, RelCalc, Draw, Energ sowie der SGML-basierten Markup-Sprache für Graphenbeschreibung NetML.
Hauptziele bei der Gestaltung von Pajek waren:

- die Unterstützung von Abstraktionen durch eine *(rekursive) Zerlegung* großer Netzwerke in mehrere kleinere Netzwerke, die mit verfeinerteren Methoden weiterbearbeitet werden können;
- die Bereitstellung einiger leistungsstarker Visualisierungswerkzeuge;
- die Implementierung einer Reihe *effizienter (subquadratischer) Algorithmen* zur Analyse großer Netzwerke.

Mit Pajek kann man: in einem Netzwerk Cluster (Komponenten, Nachbarschaften „wichtiger" Knoten, Cores usw.) auffinden, zum selben Cluster gehörige Knoten extrahieren und sie separat, auch mit Teilen der Umgebung (detailierte Lokalansicht), darstellen, Knoten in Clustern verkleinern und die Beziehungen zwischen Clustern (Globalansicht) darstellen. Neben gewöhnlichen (gerichteten, ungerichteten, gemischten) Netzwerken unterstützt Pajek auch Two-Mode-Netzwerke (bipartite (bewertete) Graphen – Netzwerke zwischen zwei disjunkten Knotenmengen) sowie temporale Netzwerke (dynamische Graphen – sich mit der Zeit verändernde Netzwerke).

Datenstrukturen

Pajek verwendet zur Analyse und Visualisierung sechs Datentypen:



- Netzwerk (Graphen),
- Partition (Namens- oder Ordnungseigenschaften der Knoten),
- Vektor (numerische Eigenschaften der Knoten),
- Cluster (Untergruppe von Knoten),
- Permutation (Neuanordnung von Knoten, Ordnungseigenschaften), und
- Hierarchie (allgemeine Baumstruktur der Knoten).

Wir beabsichtigen, diese Liste durch die Unterstützung multipler Netzwerke und Linien-Partition zu ergänzen.
Die Leistungsfähigkeit von Pajek beruht auf mehreren Transformationen, die verschiedene Übergänge zwischen diesen Datenstrukturen unterstützen. Darauf ist auch die Menüstruktur des Hauptfensters von Pajek aufgebaut. Das Hauptfenster benutzt ein „Kalkulator"-Paradigma mit einem Listengenerator für jeden Datentyp. Auch die an den jeweils aktiven (ausgewählten) Daten vollzogenen Operationen geben die Ergebnisse mithilfe von Berichtgeneratoren wieder.
Die Prozeduren sind über die Menüs im Hauptfenster aufrufbar. Häufig verwendete Abfolgen von Operationen lassen sich als Makros definieren. Damit kann Pajek auch von Usergruppen aus unterschiedlichen Bereichen (Soziale Netzwerke, Chemie, Genealogie, Computerwissenschaft, Mathematik …) für ihre Zwecke adaptiert werden. Ferner unterstützt Pajek auch wiederholte, auf Netzwerkserien angewandte Operationen.

Algorithmen

Zur Unterstützung dieser Designziele implementierten wir mehrere aus der Literatur bekannte Algorithmen, für manche Zwecke mussten wir allerdings auch neue, effiziente, für große Netzwerke geeignete Algorithmen entwickeln. Sie bieten vor allem unterschiedliche Möglichkeiten zur Identifizierung interessanter Strukturen in einem bestimmten Netzwerk. Um die Kapazitäten von Pajek zu erweitern, werden bei sehr großen Netzwerken die meisten Grundoperationen in situ ausgeführt (wobei das Input-Netzwerk zerstört wird).
In Pajek sind verschiedene bekannt effiziente Algorithmen implementiert , so etwa:

- Vereinfachungen und Transformationen: Löschen von Loops, multiple Kanten, Umwandlung von Bögen in Kanten usw.;
- Komponenten: stark, schwach, zusammenhängend, symmetrisch;
- Zerlegung: symmetrisch-azyklisch, hierarchisches Clustering;
- Pfade: kürzeste(r) Pfad(e), alle Pfade zwischen zwei Knoten;
- Flüsse: Maximalfluss zwischen zwei ausgewählten Knoten;
- Nachbarschaft: k-Nachbarn;
- CPM – kritische Pfade;
- Soziale Netzwerkalgorithmen: Zentralitätsmessung (vgl. Abb. 1), Hubs und Autoritäten, Statusmessung, Brokerrollen, strukturelle Löcher, Diffusion-Partition;
- Messung von Abhängigkeiten zwischen Partitionen / Vektoren: Cramers V, Spearman Status-Korrelationskoeffizient, Pearson-Korrelationskoeffizient, Rajski-Koeffizient;
- Extraktion von Unternetzwerken;
- Verkleinerung von Clustern im Netzwerk (allgemeine Blockmodellierung);
- Neuanordnung: topologische Anordnung, Richards-Gleichung, Murtaghs Algorithmen für Seriation und Klumpung, Depth/Breadth-First-Search.

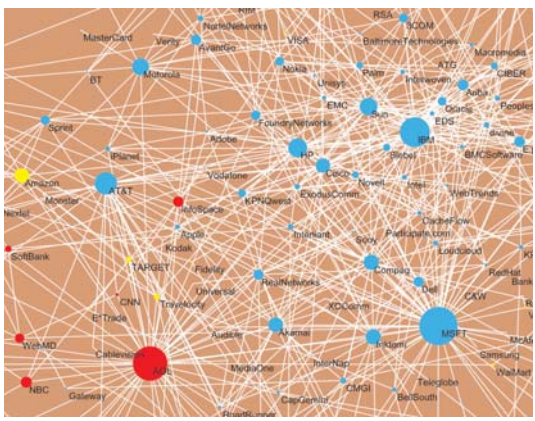


Abb. 1: Eine Zoom-Ansicht des Hauptteils der Internet-Unternehmen (gesammelt von Valdis Krebs). 219 Knoten, 631 Kanten. Jeder Netzwerkknoten repräsentiert einen Wettbewerbsteilnehmer in der Internetindustrie, 1998 bis 2001; rot – Inhalt, blau – Infrastruktur, gelb – Handel. Zwei Firmen sind mit einer Kante verbunden, wenn sie ein Joint Venture, eine strategische Zusammenarbeit oder sonstige Partnerschaft bekannt gegeben haben. Die Knotengröße entspricht der Betweenness-Zentralität.

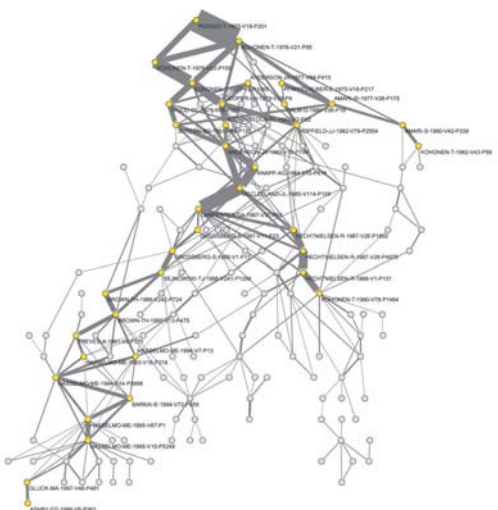


Abb. 2: Wichtigstes Unternetzwerk auf Ebene 0.007 des SOM (Selforganizing Maps) Zitiernetzwerks (4470 Knoten, 12731 Bögen). Das Bogengewicht entspricht der Anzahl der durch den Bogen verlaufenden Pfade von der Quelle zur Senke.

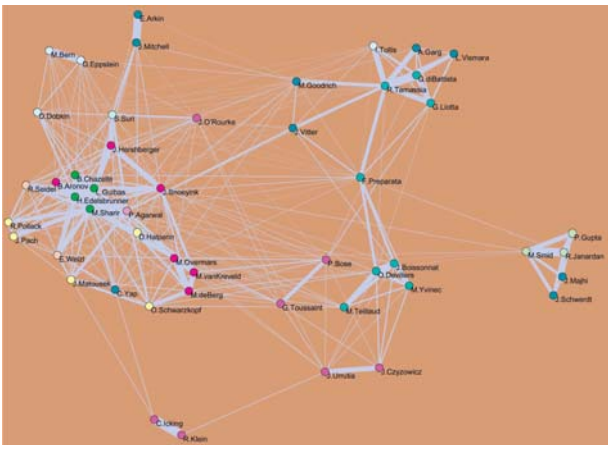


Abb. 3: pS -Core auf Ebene 46 des Kooperationsnetzwerks (7343 Knoten, 11898 Kanten, das Kantengewicht gibt die Anzahl gemeinsamer Werke wieder) auf dem Gebiet der Computergeometrie.

Spezialalgorithmen

In Pajek bauten wir auch einige Algorithmen ein, die sich eigenen Forschungen zur Analyse großer Netzwerke verdanken.

- Inseln: Wenn wir einen gegebenen oder errechneten Wert von Knoten/Linien als Höhe darstellen und das Netzwerk bis zu einer bestimmten Ebene unter Wasser setzen, ergeben sich Inseln. Diese unterscheiden sich je nach der gewählten Wasserhöhe. Inseln sind eine sehr allgemeine und effiziente Methode, um in einem bestimmten Netzwerk die „wichtigen" Unternetzwerke festzustellen.
- Zitationsgewichte: Die Analyse von Zitiernetzwerken begann 1964 mit dem Aufsatz von Garfield et al. 1989 schlugen Hummon und Doreian drei Indizes vor – Bogengewichte, mit denen wir automatisch feststellen könne, was der wichtig(st)e Teil eines Zitiernetzwerks ist. Für zwei dieser Indizes haben wir Algorithmen entwickelt, um sie effizient errechnen zu können. Vgl. Abb. 2.
- Cores und verallgemeinerte Cores: Der Begriff Core wurde 1983 von Seidman eingeführt. Knoten, die zu einem *k-Core* gehören, müssen mindestens mit *k* anderen Knoten des Core verbunden sein. Zur Bestimmung von Cores gibt es einen sehr effizienten Algorithmus. Der Begriff Core kann aber auch auf andere Knotenfunktionen ausgedeht werden, und für mehrere davon lassen sich die entsprechenden Cores effizient bestimmen. (Vgl. Abb. 3.)
- Mustersuche: Kommt ein ausgewähltes, durch einen bestimmten Graphen beschriebenes Muster in einem dünnen Netzwerk nicht oft vor, findet der zur Mustersuche eingesetzte einfache Backtrack-Algorithmus alle Vorkommen des Musters sehr rasch auf, selbst im Fall besonders großer Netzwerke. Die Mustersuche wurde erfolgreich zum Auffinden von Atomanordnungen in Molekülen (Kohlenstoffringen) und von mehrfach verbindenden Ehen in Genealogien eingesetzt.
- Triaden: Eine Triade ist ein Untergraph mit 3 Knoten. Es gibt 16 Arten von Triaden. Einige Eigenschaften eines Netzwerks lassen sich in Form seines Triadenspektrums ausdrücken – die Verteilung aller seiner Triaden.

- Dreiecksnetzwerke: Wir können einem bestimmten Graphen ein Dreiecksnetzwerk zuordnen, in dem wir jede Linie des ursprünglichen Graphen durch die Anzahl der in ihr enthaltenen Dreiecke gewichten. Die Dreiecksgewichte bilden zusammen mit den Inseln eine sehr effiziente Methode zur Feststellung der dichten Teile eines Graphen.
- Generierung großer Zufallsnetzwerke: Pajek verfügt über sehr effiziente Algorithmen zur Generierung von Zufallsnetzwerken des Erdös-Renyi-Typs (ungerichtet, gerichtet, azyklisch, ungerichtet bipartit, gerichtet bipartit, azyklisch bipartit, Two-Mode usw.). Es verfügt auch über einige Prozeduren zur Generierung zufälliger skalenfreier Zufallsnetzwerke.
- Normalisierungen: Die Normalisierungsfunktion wurde zur raschen Überprüfung von aus Two-Mode-Netzwerken abgeleiteten (One-Mode) Netzwerken entwickelt – eine Art netzwerkbasiertes Datenmining. In Netzwerken, die von großen Two-Mode-Netzwerken abgeleitet sind, gibt es oft große Gewichtsunterschiede, wodurch es nicht möglich ist, die Knoten aufgrund der Rohdaten zu vergleichen. Um die Gewichte vergleichen zu können, müssen wir das Netzwerk zuerst normalisieren. Dazu gibt es verschiedene Möglichkeiten. Zum Beispiel:

$$\mathrm{Geo}_{uv} = \frac{w_{uv}}{\sqrt{w_{uu}w_{vv}}}$$

Nach einer ausgewählten Normalisierung erhält man die wichtigen Teile des Netzwerks, indem man das normalisierte Netzwerk auf der gewählten Ebene *t* mit einer Linie schneidet und die Komponenten mit wenigstens *k* Knoten beibehält.

Algorithmen für kleine Netzwerke

Obwohl hauptsächlich zur Analyse großer Netzwerke entwickelt, wird Pajek auch oft speziell zur Visualisierung kleiner Netzwerke verwendet. Das Programm verfügt über einige Datenanalyseprozeduren, die Zeitkomplexitäten höherer Ordnung aufweisen und daher nur auf kleinere Netzwerke oder ausgewählte Teile größerer Netzwerke anwendbar sind: hierarchisches Clustering, verallgemeinerte Blockmodellierung, Partionierung markierter Graphen, TSP (Traveling Salesman Problem), Berechnung geodätischer Matrizen usw.

Layout-Algorithmen und Layout-Merkmale

Da große Netzwerke nicht detailliert in einer Gesamtansicht visualisiert werden können, müssen wir zuerst interessante Unterstrukturen eines solchen Netzwerks feststellen und sie dann in Einzelansichten visualisieren. Besonderen Wert legt Pajek auf die automatische Generierung von Netzwerklayouts. Mehrere Standardalgorithmen zur automatischen Graphenzeichnung sind

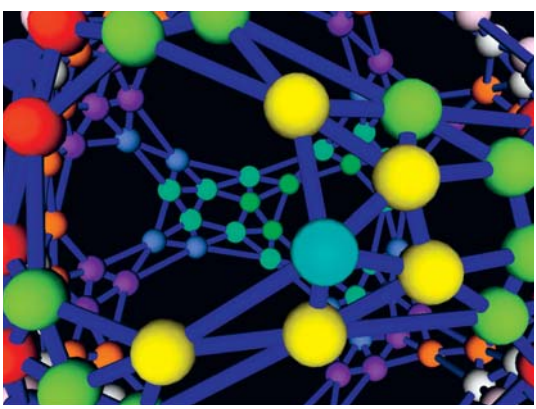


Abb. 4: Ein eigenvektor-basiertes 3D-Layout eines 5-regulären Graphen.

implementiert: Spring-Embedders (Kamada-Kawai und Fruchterman-Reingold), Layouts auf Basis von Eigenvektoren (Lanczos-Algorithmus), Zeichnen mit Zeichenebenen (Genealogien und andere azyklische Strukturen), Weitwinkelansichten und Block-(Matrix-)Darstellungen. (Vgl. Abb. 4.)
Diese Algorithmen wurden modifiziert und erweitert, um zusätzliche Optionen zu ermöglichen: Zeichnen mit Constraints (Optimierung des ausgewählten Netzwerkteils, Festsetzen einiger Knoten an vordefinierten Positionen, Verwendung von Kantenwerten als Ähnlichkeiten oder Unähnlichkeiten), Zeichnen in 3D. Pajek verfügt auch über Werkzeuge zur manuellen Bearbeitung des Graphenlayouts. Die Werte der Vektoren können zur Festlegung mehrerer Elemente der Netzwerkdarstellung, wie etwa der X-, Y-, Z-Koordinaten und der Größe der Knotenform, verwendet werden. Die Partition lässt sich graphisch durch Farbe und Form der Knoten darstellen. Auch die Werte der Kanten sind durch Stärke und/oder Farbe darstellbar. Pajek unterstützt in seinem Zeichenfenster auch das Zeichnen von Netzwerksequenzen und deren Export in Formate, die mit bestimmten 2D- oder 3D-Viewern (z. B. SVG und Marge) betrachtet werden können. Bilder in SVG können mit einem in Javascript geschriebenen Supportprogramm weiter kontrolliert werden.

Interfaces

Pajek unterstützt auch einige nicht-native Inputformate: UCINET-DL-Dateien; chemische MDLMOL- und BS-Daten; sowie genealogische GEDCOM-Daten.
Die Layouts können in folgende, von bestimmten 2- oder 3D-Viewern lesbare Grafikformate exportiert werden: Encapsulated PostScript (EPS), Scalable Vector Graphics (SVG), VRML, MDLMOL/ chime und Kinemages (Mage).
Das Menü *Tools* im Hauptfenster ermöglicht den Datenexport in die Statistikprogramme R und SPSS. Im Menü *Tools* kann der Nutzer Aufrufe seiner bevorzugten Viewer und anderer Werkzeuge vorbereiten. Es ist auch möglich, Pajek (und Makros) von anderen Programmen aus (wie R, UCINET etc.) laufen zu lassen.

Aus dem Englischen von Wilfried Prantner

Diese Präsentation von Pajek ist eine gekürzte und aktualisierte Version des Kapitels V. Batagelj, A. Mrvar: „Pajek—Analysis and Visualization of Large Networks“, in: Jünger, M., Mutzel, P. (Eds.), *Graph Drawing Software,* Berlin: Springer 2003, S. 77-103. Die Arbeit wurde in Teilen vom slowenischen Ministerium für Bildung, Wissenschaft und Sport gefördert, Projekte J1-8532 und Z5-3350.