Budapesti Műszaki és Gazdaságtudományi Egyetem
Méréstechnika és Információs rendszerek Tanszék

Rendszerbiológia – hálózati medicina

Dr. Hullám Gábor

Rendszerbiológia

- Egyes entitások (pl.gének, fehérjék) vizsgálata helyett komplex kapcsolatok és interakciós mintázatok leírása

- Heterogén, különböző omikai szinteken létező biológiai adatok állnak rendelkezésre nagy mennyiségében

- Matematikai keret: hálózatelmélet (gráfelmélet)
  - diszkrét entitások közötti kapcsolatok, mintázatok
  - hálózatok emergens tulajdonságai

Hálózatok 4 fogalmi szintje

1.) Hasonlósági hálózatok

- egyszerűen generálhatók tetszőleges hasonlósági mátrixokból.
- pl.: szekvencia hasonlósági hálózatok

2.) Leíró gráfok

- a hálózatbiológia főáramát képviselik
- pl.: a fehérje–fehérje interakciós hálózatok

3.) Függetlenségi térképek és oksági diagramok

- bioinformatika területén népszerű
- inkább statisztikai megközelítés, mint a hálózatelmélet
- pl. : Bayes-hálók

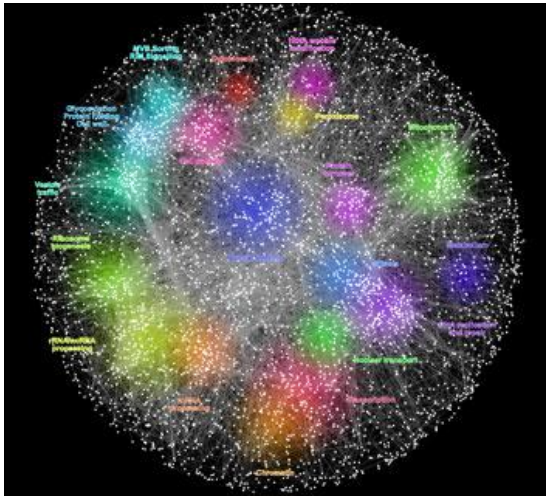
4.) Kvantitatív szabályozási hálózatok

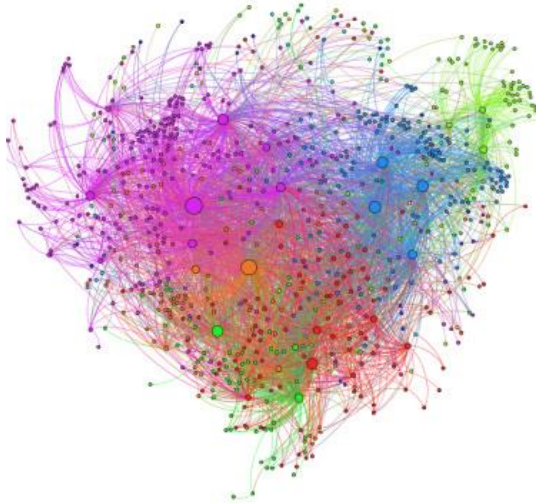
- különböző sejtszintű folyamatokat és funkciókat leíró kifinomult matematikai modellek
- gyakran közönséges és parciális differenciálegyenletek segítségével modelleznek biokémiai reakciókat

Biológiai hálózatok

Szekvencia/szerkezeti hasonlósági hálózatok

- Entitáspárokra értelmezett hasonlóságmérték meghatározásával származtatható.
- Leggyakrabban: gének, fehérjék, kismolekulák (pl. gyógyszerek)
- Esetleg elvontabb objektumok pl.: betegségek, génexpressziós profilok
- Széles alkalmazási terület : funkció és interakciók predikciója, gyógyszerkutatás



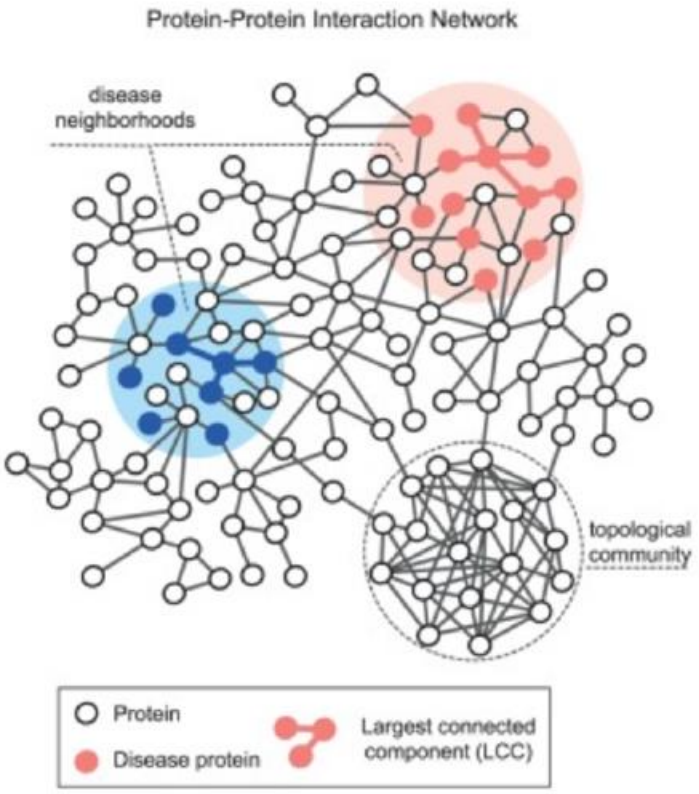


https://cherrylab.stanford.edu/blog/exploring-global-yeast-genetic-interaction-network
http://warunika.weebly.com/research.html

Biológiai hálózatok

Fehérje-fehérje interakciós hálózatok (PPI, PIN)

- építése fizikai fehérjekötődési adatok alapján történik (rendszerint nagy áteresztőképességű eszközök felhasználásával)
- Elsődleges alkalmazási terület: a fehérjék funkciójának meghatározása interakcióik elemzésével
- Publikus adatbázisok: DIP (Xenarios, et al., 2000), MINT(A. Chatr-aryamontri et al. , 2007)

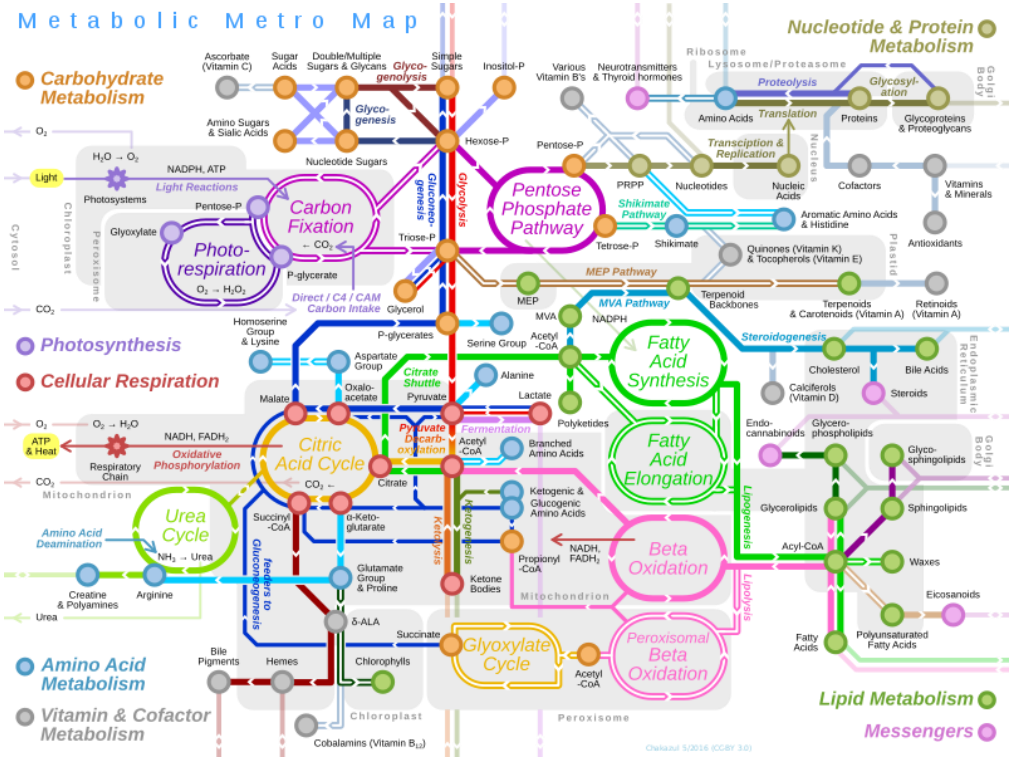


http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1004120

Biológiai hálózatok

Metabolikus hálózatok

- Építőelemei: enzimek, enzim szubsztrátok, enzimtermékek (metabolitok), a katalizált reakciók reprezentációi
- Elsődleges alkalmazási terület: élő szervezetek metabolikus útvonalainak vizsgálata
- Nyílt adatbázisok:
  KEGG (M. Kanehisa and S. Goto, 2000),
  BioCyc (R. Caspi et al.,2014)

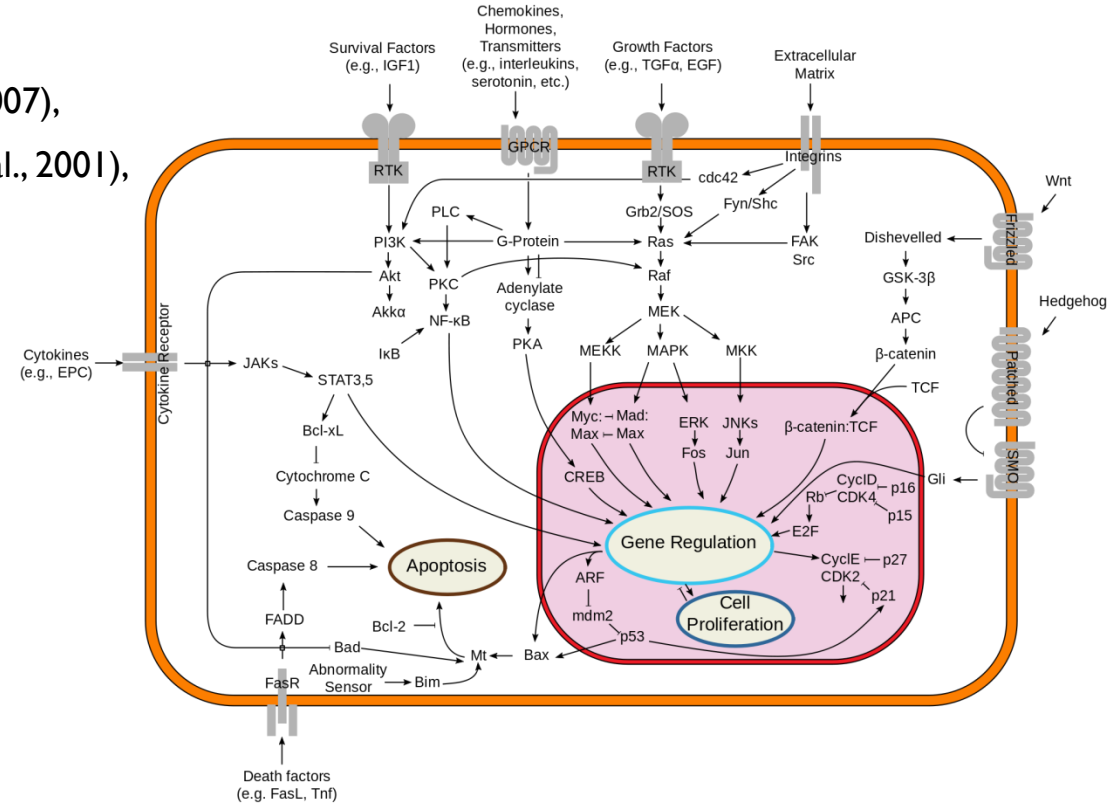


https://en.wikipedia.org/wiki/Metabolic_network

Biológiai hálózatok

Szignál transzdukciós hálózatok

- Alkalmazási terület: szignálok továbbításának vizsgálata, releváns molekuláris útvonalak és a cross-talk mechanizmusok vizsgálata.
- Elérhető adatbázisok:
  MiST (L. E. Ulrich and I. B. Zhulin, 2007),
  TRANSPATH (F. Schacherer et al., 2001),
  SignaLink (D. Fazekas et al., 2013)

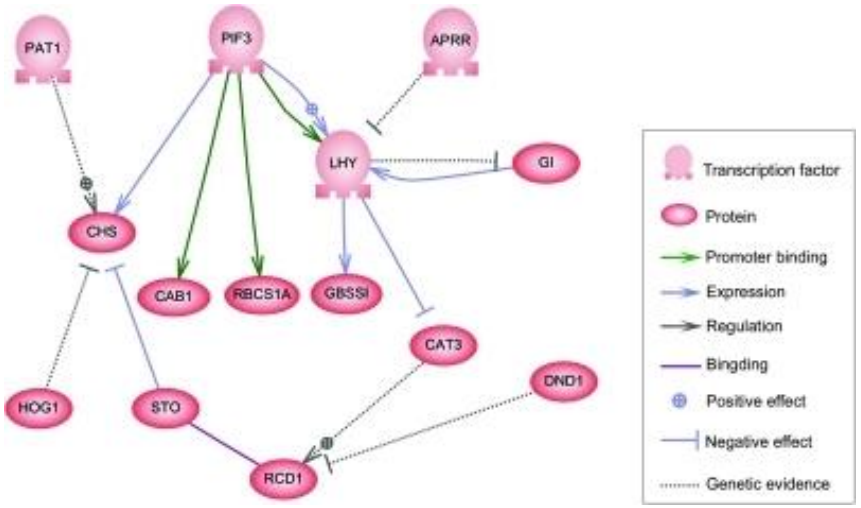


https://en.wikipedia.org/wiki/Signal_transduction

Biológiai hálózatok

Szabályozási hálózatok (GRN)

- Alkalmazási terület: a génexpresszió szabályozásának vizsgálata (szabályozási régiók, transzkripciós faktorok, RNS interferencia, poszttranszlációs módosítások, más faktorokkal történő interakciók)
- Elérhető adatbázisok:
  JASPAR (A. Sandelin et al., 2004),
  TRANSFAC (E. Wingender et al., 2000)



https://en.wikipedia.org/wiki/Gene_regulatory_network

Biológiai hálózatok

Egyéb integrált hálózatok

- Alkalmazási terület: több heterogén információforrás kombinálásával az entitások egységes nézőpontból történő vizsgálata
- Példák: többrétegű szabályozási hálózatok, gyógyszer–betegség–gén hálózatok és számos
- Connectivity Map: betegségeket, kismolekulákat és génexpressziós adatokat integrál

Gráfelméleti alapok

Gráf: egy csúcsokból és élekből álló gyűjtemény, amelyet a G = (V,E) rendezett párral jelölünk

- V a csúcsok (csomópontok) halmaza
- E az élek (vagy kapcsolatok) halmaza
  - Minden él megfeleltethető egy V -beli csúcspárnak
  - egy él mindig két, szomszédosnak nevezett csúcsot köt össze (ez a kettő lehet ugyanaz a csúcs bizonyos esetekben)

Élek irányítása

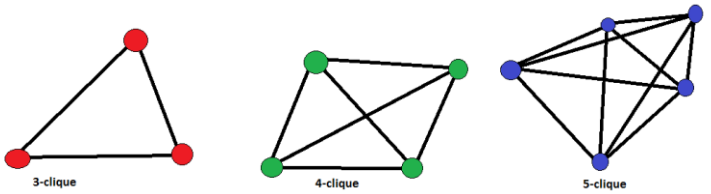
- Ha van: akkor a kérdéses gráf egy irányított gráf
  - az élek rendezett csúcspárokként reprezentálhatók (Szülő, Gyermek)
  - a kapcsolat nem szimmetrikus
  - speciális esete az irányított körmentes gráf (DAG),
- Ha nincs: akkor a kérdéses gráf egy irányítatlan gráf
  - a kapcsolat szimmetrikus

Alapfogalmak

- Súlyozott gráf – súlyozott él: az élekhez számszerű értékeket rendelünk.
- Fokszám: Egy adott csúcsra illeszkedő (kapcsolódó) élek száma
- Szabályos gráf: minden csúcs fokszáma megegyezik.
- Teljes gráf: a szabályos gráf speciális esete, ahol bármely két csúcsra illeszkedik él.
- Összefüggő gráf: ha bármely két csúcsa között létezik út – ellenkező esetben a gráf nem összefüggő.
- Részgráf: az eredeti gráf kiválasztott csúcsaiból és éleiből áll, ahol a kiválasztott élek a kiválasztott csúcsokra illeszkednek.
- Komponens: maximális (lehető legnagyobb) összefüggő részgráf
  (Egy nem összefüggő gráf több komponens tartalmaz, míg egy összefüggő gráf pontosan egyet.)

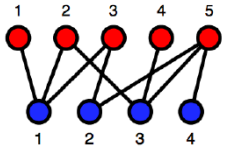
Alapfogalmak

- Klikk: Egy gráf teljes részgráfjai



- Maximális klikk: a lehető legnagyobb klikk

- Páros gráf: a csúcsok két diszjunkt halmazt alkotnak, ahol azonos halmazbeli csúcsokra nem illeszkedik él



- Klaszter: a csúcshalmaz egy olyan részhalmaza, amelyben a csúcsok „sokkal erősebben” kapcsolódnak egymáshoz, mint a gráf többi részéhez.

- További fontos mértékek: a legrövidebb út, az átlagos úthossz, a hálózati centralizáció, csomóponti centralitások

Hálózatelemzés

A hálózatelemzés a hálózat kvalitatív és kvantitatív tulajdonságait vizsgálja:

- mögöttes strukturális alapelvek
- funkcionális szerveződés
- lokális mintázatok
- emergens tulajdonságok
- dinamikus viselkedés

- Hasonló eszközöket használnak telekommunikációban, szociális hálózatok elemzésében és számos egyéb területen

Hálózati topológia

A hálózati topológia: a csomópontok és kapcsolataik elrendeződését jellemzi

- A gráfok gyakran rendelkeznek jól meghatározott strukturális elemekkel ezek egy része jelentősen befolyásolja a hálózat viselkedését:

Lokális topológiai struktúrák:

Hub: Az átlagosnál sokkal több kapcsolattal rendelkező csomópont

- a hálózat kulcsszereplői: törlésük rendszerint a hálózat gyors degradációjához, izolált klaszterekre való széteséséhez vezet.
- Ez a jelenség PPI hálózatok esetén „centralitási–letalitási szabály” (hub-ok ~ nélkülözhetetlen fehérjék)

Motívum: szignifikánsan felülreprezentált irányított részgráfok

Graphlet: szignifikánsan felülreprezentált irányítatlan részgráfok

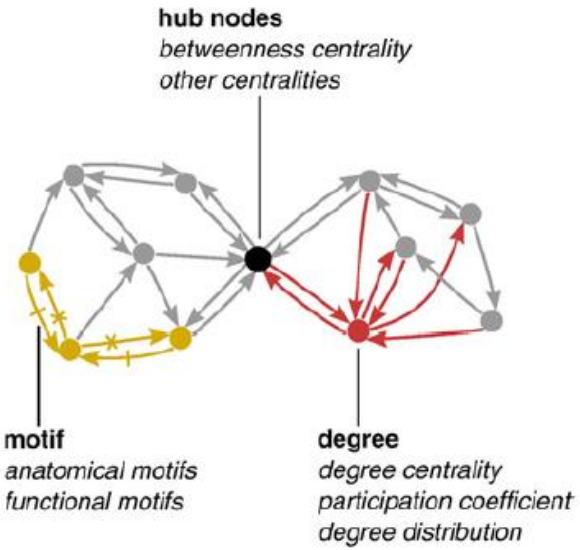
Hálózati topológia

Hub: Az átlagosnál sokkal több kapcsolattal rendelkező csomópont

- a hálózat kulcsszereplői: törlésük rendszerint a hálózat gyors degradációjához, izolált klaszterekre való széteséséhez vezet.
- Ez a jelenség PPI hálózatok esetén „centralitási–letalitási szabály” (hub-ok ~ nélkülözhetetlen fehérjék)

Motívum: szignifikánsan felülreprezentált irányított részgráfok

Graphlet: szignifikánsan felülreprezentált irányítatlan részgráfok



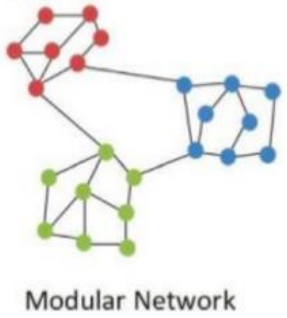
Hálózati topológia

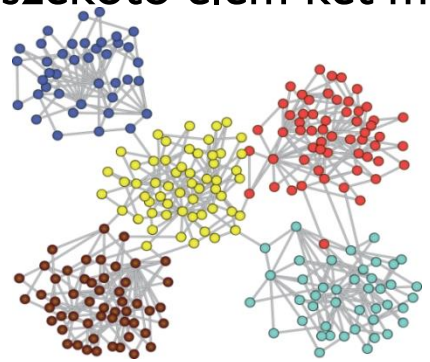
Modul: többé-kevésbé a gráfelméleti klaszternek felel meg.

- Gyakran funkcionális alrendszereket reprezentálnak, pl. bizonyos sejtszintű folyamatokat vagy funkciókat.
- Összetett rendszerekben több típusú interakció is elképzelhető az egyes modulok között, például átlapolódáson vagy hidakon keresztül
- A modulok hierarchikus elrendeződést is mutathatnak; kisebb, interakcióban lévő modulok nagyobb, lazább modulok alkotóiként szerepelhetnek

Híd: modulokat összekötő csomópontok

Bottleneck: Ha egy híd az egyetlen összekötő elem két modul között



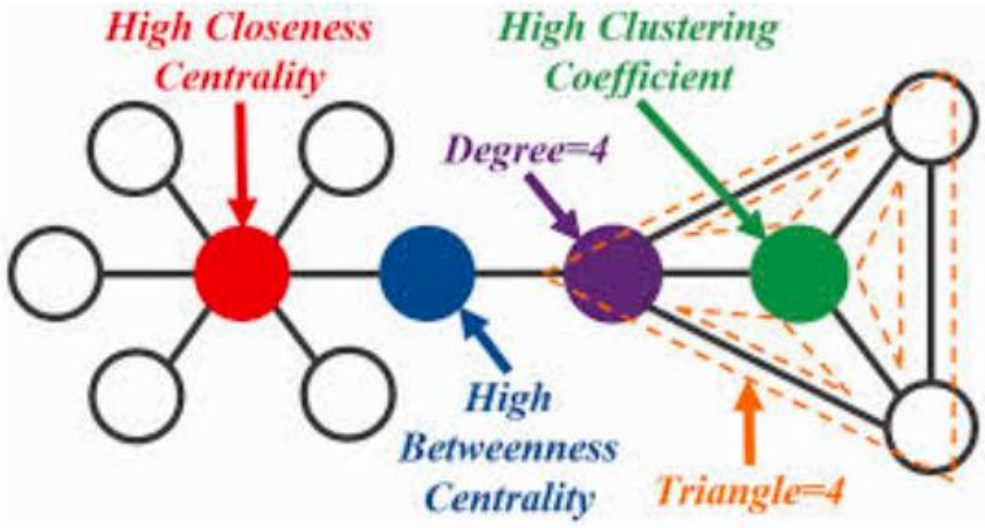


https://openi.nlm.nih.gov/detailedresult.php?img=PMC2542420_pcbi.1000190.g004&req=4

Hálózati topológia

Csomóponti centralitás: általánosságban a csomópontok „befolyásosságának mértéke”

- Pl.: Ha léteznek a hálózat egyfajta globális „koordinátoraként” viselkedő csomópontok, akkor ezek magas centralitással bírnak



Hálózati topológia

Hálózati centralizáció: a csomóponti centralitások eloszlását írja le

- Erősen centralizált hálózatok gyakran csillagszerű topológiát mutatnak
- A magas centralitású csomópontokból álló alhálózatot csontváznak nevezzük.
- Kis világ tulajdonság: alacsony átlagos úthossz, a hálózat esetenként hatalmas mérete ellenére

Hálózati modellek és dinamika

- A valóságban sok hálózat időben folyamatosan változik és fejlődik.

Hálózat dinamika: temporális aspektusokat hivatott vizsgálni.

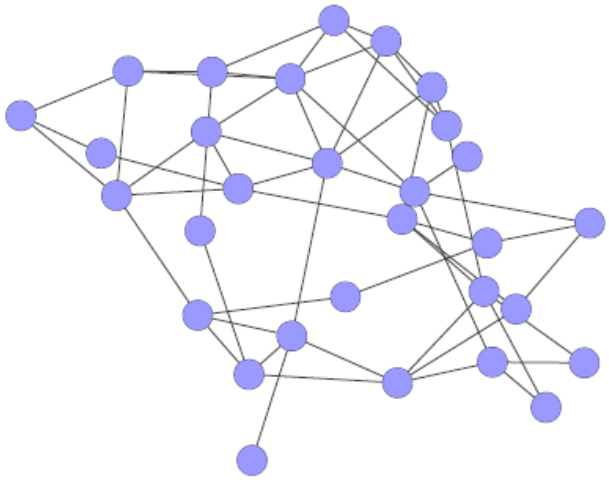
- Célja feltárni, hogyan következnek az emergens tulajdonságok a kis számú, egyszerű konstrukciós szabályból.

Híres hálózati modellek:

- Erdős–Rényi-modell
- Watts–Strogatz-modell
- Barabási–Albert-modell

Erdős–Rényi modell

- Az egyik legegyszerűbb modell véletlen gráfok leírására
- A konstrukció N csomóponttal indul, majd véletlenszerűen húz be éleket az N(N - 1)/2 lehetőségből
- Rendelkezik a kisvilág-tulajdonsággal
- A fokszámok között csak kis variancia tapasztalható, azaz nem képesek megmagyarázni a valós hálózatok klasztereződési tendenciáját (pl. hubok formálódását)

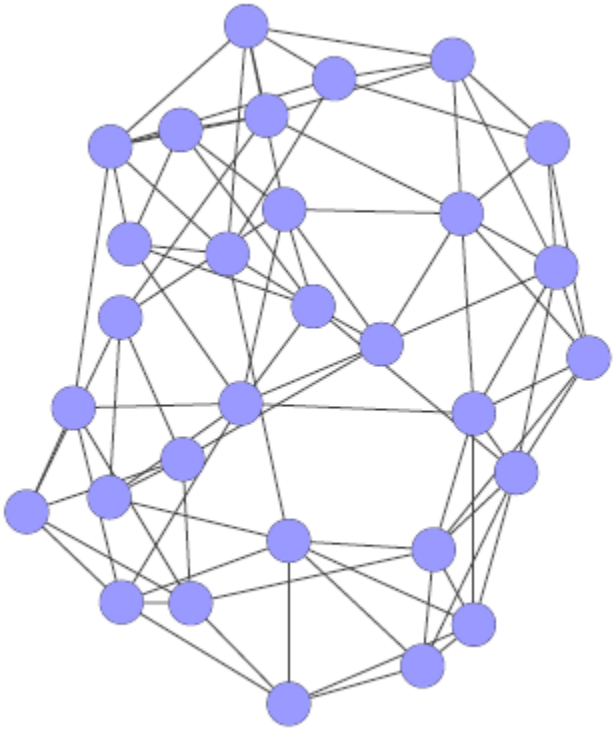


Watts–Strogatz – modell

- mind a kisvilág-tulajdonságot, mind a lokális klasztereződést reprodukálja
- Konstrukció:
  - kezdetben az N darab csomópont egy körben van elrendezve, továbbá minden csomópont össze van kötve k/2 legközelebbi szomszédjával.
  - Ezután minden él egy kis p valószínűséggel „áthuzalozódik”, azaz egyik vége egy véletlenszerűen kiválasztott csomóponthoz csatlakozik – ennek köszönhető a kisvilág-tulajdonság.

Watts–Strogatz – modell

- Ha p-t megfelelően, de nem extrém módon kicsire választjuk, elfogadható mértékű lokális klasztereződés marad a hálózatban

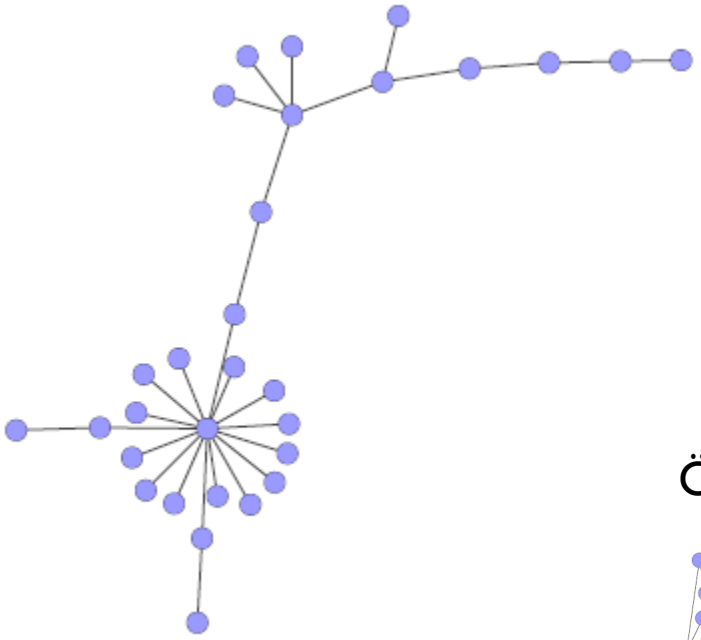


Barabási–Albert-modell

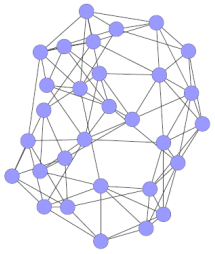
- Mind a kisvilág-tulajdonságot, mind a lokális klaszterződést reprodukálja
- Skálafüggetlen fokszámeloszlást mutat, amely gyakran megfigyelhető valós hálózatokban, például a biológia területén vagy az Interneten
- Konstrukció: A modell alapötlete a növekedés és preferenciális kapcsolódás alkalmazása.
- Növekedés: Új csomóponttal való kiegészülés
  - A többi csomópont aktuális fokszámát figyelembe véve alakulnak ki az új csomópont kapcsolatai, azaz az új csomópont a már eddig is sok kapcsolattal rendelkezőket preferálja a kapcsolódás során

Barabási–Albert-modell

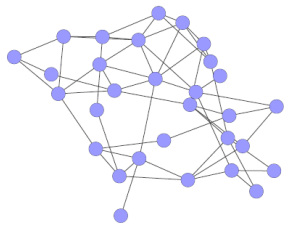
- A preferenciális kapcsolódás hűen modellezi számos valós (pl. szociális) hálózat formálódási szabályait



Összehasonlításképp:



Watts-Strogatz



Erdős-Rényi

Asszortativitás

- Asszortativitás: a csomópontok „hasonló” csomópontokhoz történő preferenciális kapcsolódását írja le
  - „hasonló” alatt rendszerint hasonló fokszámot értünk.

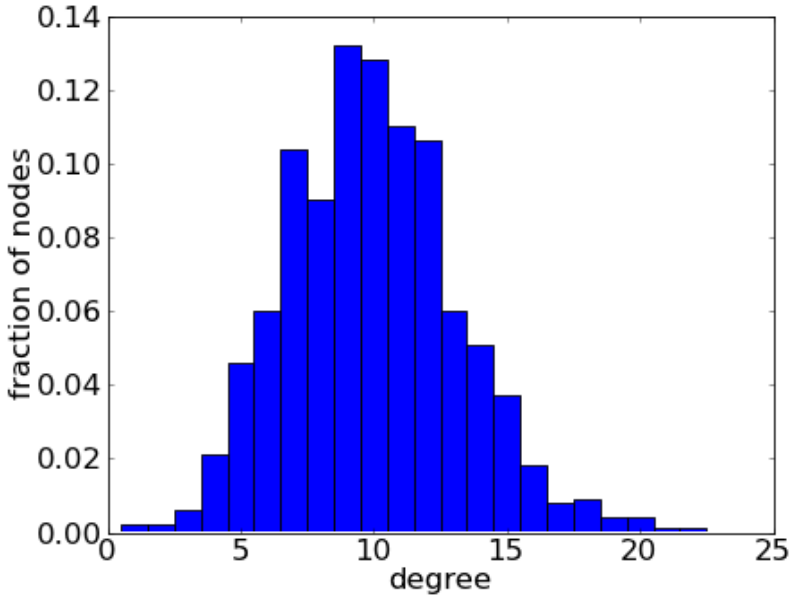
- Asszortatív hálózatokban a sok kapcsolattal rendelkező csomópontok más, sok kapcsolattal rendelkező csomópontokat preferálnak

- A biológiai hálózatok rendszerint diszasszortatívek, azaz magas fokszámú csomópontok alacsony fokszámúakhoz kapcsolódnak

Fokszámeloszlás és skálafüggetlen hálózatok

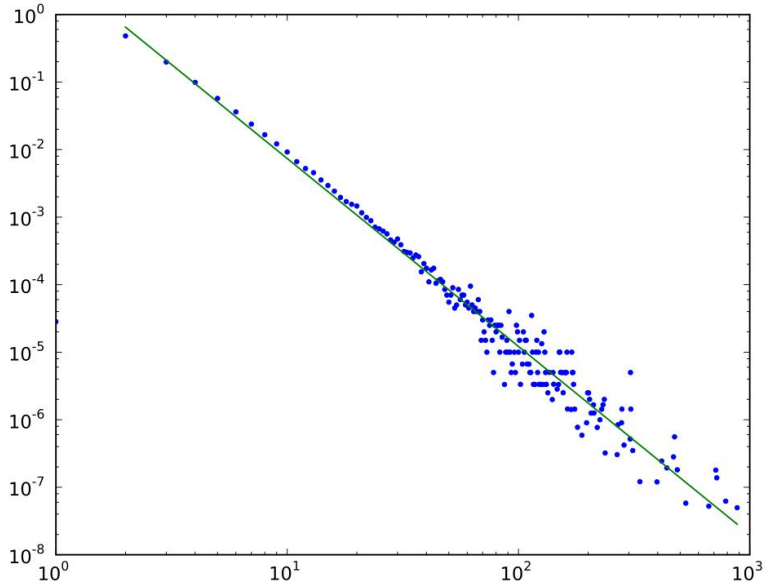
- Fokszámeloszlás: (p(k)) annak valószínűségét adja meg, hogy egy csomópont fokszáma pontosan k.
- A biológiai hálózatok további kulcsfontosságú tulajdonsága, hogy a fokszámeloszlás hatványfüggvényt követ, ún. skálafüggetlen hálózatot eredményezve.

- Erdős–Rényi-modellben a fokszámeloszlás binomiális, ami nagy (átlagostól nagyon eltérő fokszámú csomópontok extrém ritkák).



Fokszámeloszlás és skálafüggetlen hálózatok

- A skálafüggetlen hálózatok $p(k) \sim k^{-\gamma}$ alakú fokszámeloszlást követnek, így néhány magas fokszámú csomópontra (hubok) sok alacsony fokszámú jut
- A fokszámkitevő alapvetően meghatározza a hálózat viselkedését.
  - $\gamma > 3$ értékeknél nagy hubok már csak elvétve fordulnak elo és nem játszanak lényeges szerepet
  - $\gamma \leq 3$ alacsonyabb értékeinél a hubok jelenléte kifejezett
  - A legtöbb biológiai hálózat fokszámkitevője 2 és 3 között van.



Feladatok - kihívások

- Csomópontok és kapcsolatok jóslása az egyik legkézenfekvőbb feladat. Csomópontok és kapcsolatok jósolhatók például hasonlóságok, topológiai vagy temporális tulajdonságok, vagy hálózati összehasonlítás felhasználásával.

- Klaszteranalízis használható funkcionális modulok elismerésére és interakcióik elemzésére biológiai rendszerekben.

- Centralitás-elemzés, útkeresés, robosztusság elemzése használható a hálózat szerveződésének megértésére és a csomópontok „kommunikációjának” leírására.
  - Pl.: gyógyszercélpontok azonosítása, azaz annak eldöntése, hogy milyen csomópontokat vagy éleket érdemes megtámadni egy betegség hatásainak kiküszöbölése érdekében

Feladatok - kihívások

- Gráf-izomorfizmus és hálózatillesztés, motívumkeresés
  - Pl.: több faj PPI hálózatainak illesztésével a fehérjék funkcionális ortológiájára következtethetünk
- Hálózatok becslése vagy „visszafejtése” (reverse engineering) : A hálózat struktúrájának adatokból történő meghatározása
- Hálózat-integráció: célja több hálózat kombinációja (tudásfúzió)
- Hálózat-vizualizáció: a legegyszerűbb, mégis a legfontosabb feladatok egyike
  - Pl.: Cytoscape rendszer valószínűleg a legnépszerűbb eszköz biológiai hálózatok vizualizációjára;

Hálózatok statisztikai értelmezése

Több megközelítés lehetséges:

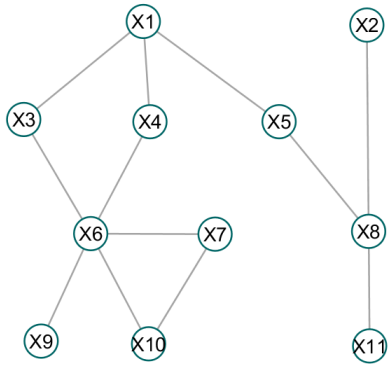
- Oksági (kauzális) modell
- Prediktív modell
- Egymással interakcióra lépő komponensek modellje
- Információ terjedésének modellezése

Oksági hálózatok

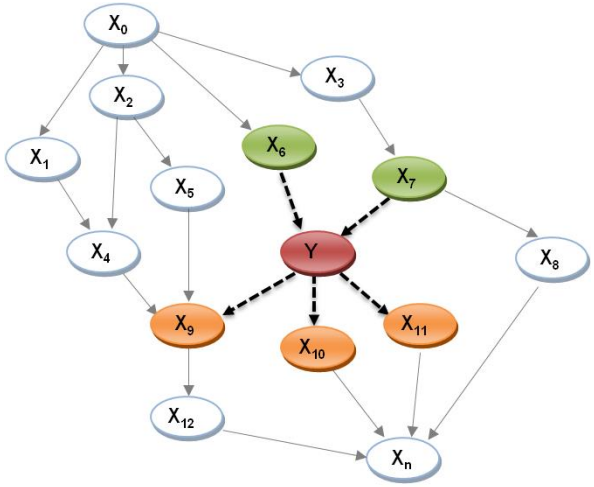
- Szakértői tudás alapján kialakított oksági struktúrák tesztelésére léteznek módszerek
  - strukturális egyenlet modellezés (Structural Equation Modeling – SEM)
- Oksági struktúrák feltárása, tanulása nehéz feladat
  - Egyértelmű feltáráshoz beavatkozásos kísérletek sorozata szükséges
  - Megfigyelések alapján csak bizonyos struktúrák tanulhatók, teljes struktúra többnyire nem
- A legtöbb modellnél feltételezzük, hogy nincsenek körkörös ok-okozati hurkok
- Reprezentáció: irányított aciklikus gráf alapú
  - SEM, oksági Bayes-hálók

Prediktív hálózatok

- Prediktív hálózatokban A és B csomópont között akkor fut él, ha A alapján prediktálható (jósolható) B
- Reprezentáció lehet:
  - irányítatlan gráf alapú



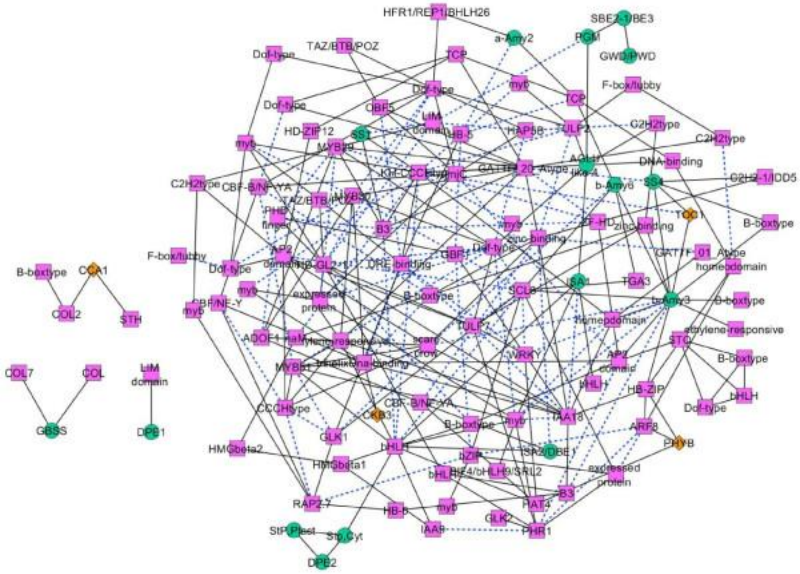
  - irányított körmentes gráf alapú (directed acyclic graph –DAG)



- Oksági modellek megfeleltethetőek prediktív modelleknek, de fordítva nem

Asszociációs hálózatok

- A és B csomópont között akkor fut él, ha A alapján prediktálható (jósolható) B függetlenül minden más csomóponttól
- Ekkor az élek valamilyen hatáserősség mértéket fejeznek ki pl.: korreláció vagy esélyhányados
- Egyszerűen kialakítható, irányítatlan gráf
- Sok olyan kapcsolat lehet benne, ami valójában nem közvetlen
- Korlátozottan alkalmas hálózatelemzéshez

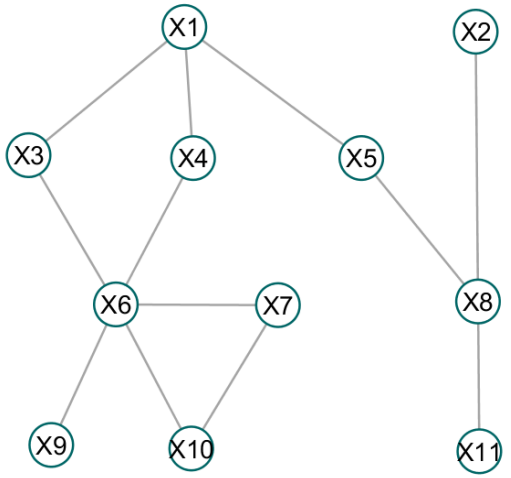


https://www.researchgate.net/figure/230684929_fig2_The-gene-association-network-of-starch-metabolism-inferred-from-GGM-Q005-The

Modellezés irányítatlan gráfokkal

Adott: *k* megmért változó *N* mintán (megfigyelésen)

Cél: Olyan nem irányított gráf kialakítása, amely a változók közötti függőségeket reprezentálja



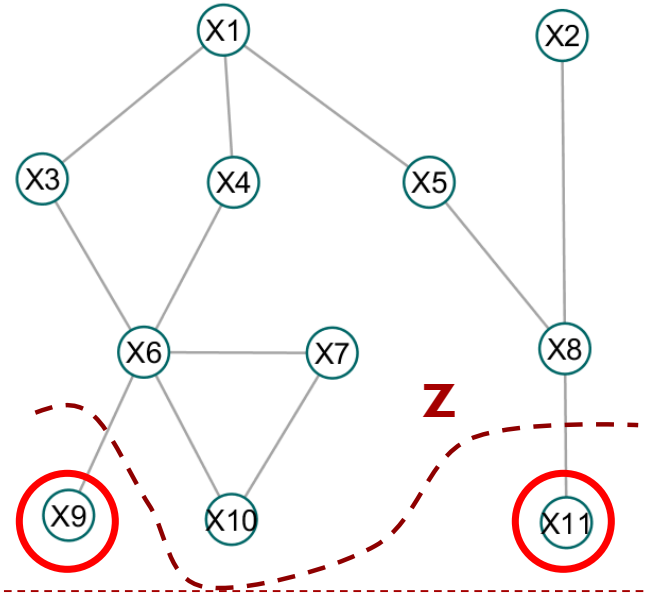
- A gráf elemei:
  - Csúcsok (Nodes) : változók
  - Élek (Edges): változók közötti feltételes függőségi kapcsolat

Feltételes függetlenség

- Hiányzó él jelentése: a két változó feltételesen független egymástól a többi változót feltéve.
- A változó feltételesen független B-től, feltéve C-t ha P(A|B,C) =P(A|C)

Példa:

- Legyen $\mathbf{V}=\{X_1,X_2,\ldots,X_{11}\}$,
  $\mathbf{Z} = \mathbf{V} \setminus \{X_9,X_{11}\}$,
  $P(X_9|X_{11},\mathbf{Z})=P(X_9|\mathbf{Z})$, akkor:
  $X_9 \not\perp X_{11}$

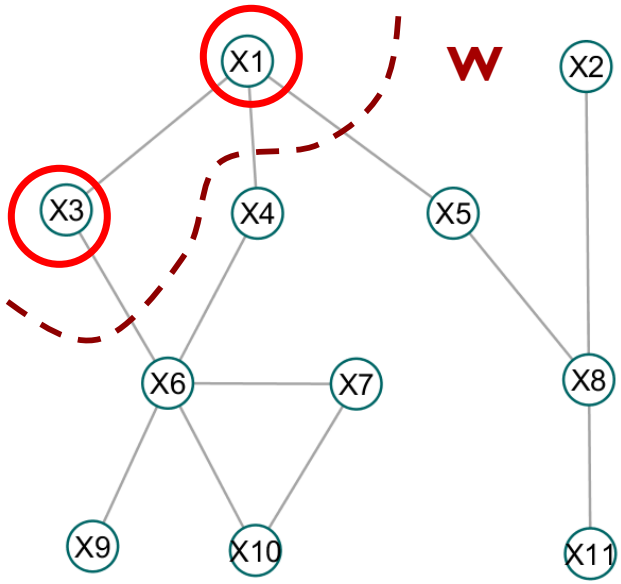


Feltételes függőség

- Létező él: a két változó között feltételes függőség áll fenn a többi változót feltéve

Pl.: Legyen $\mathbf{W} = \mathbf{V} \setminus \{X_1, X_3\}$, és $X_1 - X_3$, ekkor $P(X_1|X_3,\mathbf{W}) \neq P(X_1|\mathbf{W})$

- Gauss eloszlást követő változóknál: ez azt jelenti, hogy
  - a köztük lévő parciális korreláció nem nulla
  - a két változó korrelált a többi változót feltéve
  - Pl.: $\rho_{X1X3.X2,X4,\ldots X11} \neq 0$



Irányítatlan gráf építése függőségek alapján

- Egy lineáris regresszió j-dik prediktorának regressziós koefficiense arányos a függő változó és a j-dik prediktor közötti parciális korrelációval

- Függő változó: Y
- Prediktor változók: $X_1, X_2, \ldots, X_k$
- Regressziós egyenlet : $Y = \beta_0 + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \ldots + \beta_k \cdot X_k$

$$Y = \beta_0 + \sum_{i=1} \beta_i \cdot X_i$$

Állítás: $\beta_i \sim \rho_{Y,Xi \,.Xj \neq i}$

Parciális korreláció - emlékeztető

- Korrelációs együttható (Pearson ρ)

$$\rho_{X,Y} = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y} \qquad \rho_{X,Y} = \frac{\operatorname{E}[XY] - \operatorname{E}[X]\operatorname{E}[Y]}{\sqrt{\operatorname{E}[X^2] - [\operatorname{E}[X]]^2}\,\sqrt{\operatorname{E}[Y^2] - [\operatorname{E}[Y]]^2}}.$$

- Parciális korreláció (rekurzív formula) $\mathbf{Z}=\{Z_0,Z_1,\ldots Z_k\}$

$$\rho_{XY\cdot\mathbf{Z}} = \frac{\rho_{XY\cdot\mathbf{Z}\setminus\{Z_0\}} - \rho_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}\rho_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}{\sqrt{1-\rho^2_{XZ_0\cdot\mathbf{Z}\setminus\{Z_0\}}}\sqrt{1-\rho^2_{Z_0Y\cdot\mathbf{Z}\setminus\{Z_0\}}}}.$$

- Parciális korreláció egy független változó esetén (Z)

$$\rho_{XY\cdot Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{ZY}}{\sqrt{1-\rho^2_{XZ}}\sqrt{1-\rho^2_{ZY}}}.$$

Naiv algoritmus - Meinshausen & Bühlmann (2006)

Alapötlet:

- Alkalmazzunk lasso regularizációt a gráf csomópontjainak megfelelő minden változóra úgy, hogy egy lépésben mindig az egyik változó a függő változó
- Az X_i és X_j csomópontok között akkor legyen él, ha
  - Az X_j prediktor változó regressziós koefficiense X_i függő változó regressziós egyenletében nem nulla VAGY
  - Az X_i prediktor változó regressziós koefficiense X_j függő változó regressziós egyenletében nem nulla

Lasso

- LASSO = **L**east **A**bsolute **S**hrinkage and **S**election **O**perator

Tibshirani R. (1994). Regression Shrinkage and Selection Via the Lasso. Journal of the Royal Statistical Society, Series B, 58, 267--288

$$\arg\min_{\beta j} \left\{ \sum_{i=1}^{N} (y_i - \sum_j x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

négyzetes hiba regularizáció

λ: regularizációs paraméter
βj:regressziós koefficiens

= L1regularizált legkisebb négyzetes hiba

Tehát a cél kettős:

- a lehető legjobban közelíteni *Y*-t a regresszióval, tehát a négyzetes hiba legyen a lehető legkisebb
- A lehető legkevesebb *Xj* segítségével, mivel a β_j koefficiensek összegével arányos a „büntetés” (regularizáció)

Lasso következménye

- Változó kiválasztás:
  - Csak azon *Xj* változók „játsszanak szerepet” a regressziós egyenletben, melyek elengedhetetlenek *Y* predikciójához
  - Ezen *Xj* változók regressziós koefficiense ne legyen 0

- Zsugorítás (shrinkage):
  - Az *Y* változó szempontjából elhanyagolható *Xi* változók koefficiense legyen 0.

Többváltozós nézőpont

- Ha $X = (X_1, X_2, \ldots, X_k)$ többváltozós Gauss eloszlást követ $\mu=0$ várhatóértékkel és Σ kovariancia mátrixszal…

$$\Sigma = \begin{bmatrix} \mathrm{cov}(X_1,X_1) & \mathrm{cov}(X_1,X_2) & \cdots & \mathrm{cov}(X_1,X_n) \\ \mathrm{E}[(X_1-\mu_1)(X_1-\mu_1)] & \mathrm{E}[(X_1-\mu_1)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_1-\mu_1)(X_n-\mu_n)] \\ \mathrm{E}[(X_2-\mu_2)(X_1-\mu_1)] & \mathrm{E}[(X_2-\mu_2)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_2-\mu_2)(X_n-\mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{cov}(X_n,X_1) & \mathrm{cov}(X_n,X_2) & \cdots & \mathrm{cov}(X_n,X_n) \\ \mathrm{E}[(X_n-\mu_n)(X_1-\mu_1)] & \mathrm{E}[(X_n-\mu_n)(X_2-\mu_2)] & \cdots & \mathrm{E}[(X_n-\mu_n)(X_n-\mu_n)] \end{bmatrix}.$$

- Átlós elem: cov(Xi,Xi)= variancia: átlagtól való eltérésnégyzet
- Nem átlós elem: cov(Xi,Xj)= két változó együttes változása

Inverz kovariancia mátrix

- Ha $X = (X_1, X_2, \ldots, X_k)$ többváltozós Gauss eloszlást követ $\mu=0$ várhatóértékkel és Σ kovariancia mátrixszal, akkor $\Theta = \Sigma^{-1}$ az inverz kovariancia mátrix (pontosság mátrix) reprezentálja minden Xj feltételes eloszlását a többihez képest

$$\Theta = \begin{pmatrix} \theta_{1,1} & \theta_{1,2} & \cdots & \theta_{1,n} \\ \theta_{2,1} & \theta_{2,2} & \cdots & \theta_{2,n} \\ \vdots & \vdots & \cdots & \vdots \\ \theta_{n,1} & \theta_{n,2} & \cdots & \theta_{n,n} \end{pmatrix}$$

- kovariancia ~ szóródás (átlaghoz képest)
- inverz kovariancia ~ koncentráció (átlaghoz képest), pontosság

- Átlós elem ($\theta_{i,i}$): átlag körüli koncentráció
- Nem átlós elem: ($\theta_{i,j}$): két változó együttes koncentrációja

Inverz kovariancia és a regressziós együtthatók kapcsolata

- Ha $\theta_{i,j}=0$, akkor X_i és X_j változók feltételesen függetlenek egymástól, feltéve a többi változót
- A változók közötti függőségeket reprezentáló irányítatlan gráf akkor tartalmaz élt X_i és X_j között, ha a két változóra vonatkoztatott inverz kovariancia nem 0.
- Gauss eloszlást (normális eloszlást) követő Xj változó eloszlásának és a regressziós egyenletének kapcsolata:

$$P(X_j \mid X_{i\neq j}) \sim N(\Sigma_{i\neq j} X_i \cdot \beta_{i,j}, \sigma_{j,j}), \text{ ahol}$$

$$\beta_{i,j} = - \theta_{i,j} / \theta_{j,j}$$

$$\sigma_{j,j} = 1/ \theta_{j,j}$$

- $\beta_{i,j} = 0 \leftrightarrow \theta_{i,j} = 0, \quad \beta_{i,j} = 0 \leftrightarrow \rho_{i,j} = 0,$

Modellillesztés – naiv megvalósítás

Naiv megvalósítás:

- illesszünk lineáris regressziót minden változóhoz külön, és hagyjuk ki azon változókat, ahonnan az adott változóhoz nem fut él
- Használjuk fel a regressziós koefficienseket a Σ^{-1} megfelelő elemeinek számításához

Probléma:

- Így az egyes változókhoz illesztett regressziók (részmodellek) függetlenek lesznek egymástól

Modellillesztés – iteratív megoldás

- 1. Inicializáljuk W-t úgy, hogy W= S (=a minta kovariancia mátrixa)
- 2. For j = 1, 2, . . . , p, 1, 2, . . . , p, . . . (ismétlés konvergencia eléréséig)
  - (a) Particionáljuk a W mátrixot 2 részre
    - 1: Összes sor és oszlop kivéve a j-edik
    - 2: A j-edik sor és oszlop

$$\begin{pmatrix} \mathbf{W}_{11} & w_{12} \\ w_{12}^T & w_{22} \end{pmatrix} \begin{pmatrix} \mathbf{\Theta}_{11} & \theta_{12} \\ \theta_{12}^T & \theta_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I} & 0 \\ 0^T & 1 \end{pmatrix}$$

  - $W_{11}\,\theta_{12} + w_{12}\,\theta_{22}=0$ $\qquad \beta = -\theta_{12} / \theta_{22}$
  - $w_{12} = -W_{11}\,\theta_{12} / \theta_{22} = W_{11}\,\beta$

Modellillesztés – iteratív megoldás

- 1. Inicializáljuk W-t úgy, hogy W= S.
- 2. For j = 1, 2, . . . , k, 1, 2, . . . , k, (ismétlés konvergencia eléréséig)
  - (a) Particionáljuk a W mátrixot 2 részre
    - 1: Összes sor és oszlop kivéve a j-edik
    - 2: A j-edik sor és oszlop
  - (b) Oldjuk meg a $W^*_{11}\beta^* - s^*_{12} = 0$ egyenletet β^* paraméterekre (redukált egyenletrendszer, a nem szükséges Xi változók kimaradnak) $\hat{\beta}$ számítása β^* alapján (feltöltés 0-ákkal a megfelelő helyeken)
  - (c) Frissítjük $w_{12} = W_{11}\hat{\beta}$
- 3. Utolsó ciklusban (minden j-re) számítsuk ki:

$$\theta_{12}\hat{} = -\hat{\beta} \cdot \theta_{22}\hat{}, \text{ az}$$

$$1/\theta_{22}\hat{} = s_{22} - w_{12}^T\hat{\beta} \text{ egyenlet alapján}$$

Modellillesztés – iteratív eljárás összegzése

$\mathbf{V} = (X_1, X_2, \ldots, X_k) \sim N(0, \Sigma)$

Többváltozós regressziós modell illesztése

- $\Theta = \Sigma^{-1}$ inverz kovariancia mátrix
- Θ log-likelihoodja:

$$\ell(\Theta) = \underbrace{\log(\det \Theta)}_{\sim Y} - \underbrace{\mathrm{tr}(S\,\Theta)}_{\sim X_i\beta_i}, \text{ ahol}$$

  - S: az adat alapján megfigyelt kovariancia mátrix
  - Det (determinant): az adott mátrix determinánsa

$$|A| = \begin{vmatrix} a & b \\ c & d \end{vmatrix} = ad - bc.$$

$$|A| = \begin{vmatrix} a & b & c \\ d & e & f \\ g & h & i \end{vmatrix} = a\begin{vmatrix} e & f \\ h & i \end{vmatrix} - b\begin{vmatrix} d & f \\ g & i \end{vmatrix} + c\begin{vmatrix} d & e \\ g & h \end{vmatrix}$$
$$= aei + bfg + cdh - ceg - bdi - afh.$$

  - tr (trace): az adott mátrix nyoma $\mathrm{tr}(A) = \sum_{i=1}^{n} a_{ii} = a_{11} + a_{22} + \cdots + a_{nn}$

Grafikus Lasso

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3), 432–441.

- $\ell(\mathbf{\Theta}) = \underbrace{\log(\det \mathbf{\Theta})}_{\sim Y} - \underbrace{\mathrm{tr}(S\,\mathbf{\Theta})}_{\sim X_i\beta_i} + \underbrace{\lambda \cdot \sum_{j \neq k} |\theta_{j,k}|}_{\text{regularizáció}}$

Gradiens egyenlet:

- $\mathbf{\Theta}^{-1} - S - \lambda \cdot \mathrm{sign}(\mathbf{\Theta}) = 0$

Modellillesztés – Grafikus lasso

- 1. Inicializáljuk W-t úgy, hogy $W = S + \lambda \cdot I$
- 2. For j = 1, 2, . . . , k, 1, 2, . . . , k,
  (ismétlés a konvergencia eléréséig)
  - (a) Particionáljuk a W mátrixot 2 részre
    - 1: Összes sor és oszlop kivéve a j-edik
    - 2: A j-edik sor és oszlop
  - (b) Oldjuk meg a $W^*_{11}\beta^* - s^*_{12} + \lambda \cdot \text{sign}(\beta^*) = 0$ egyenletet β^* paraméterekre (redukált egyenletrendszer, a nem szükséges Xi változók kimaradnak, $\beta\hat{}$ számítása β^* alapján (feltöltés 0-ákkal a megfelelő helyeken)#
  - (c) Frissítjük $w_{12} = W_{11}\beta\hat{}$
- 3. Utolsó ciklusban (minden j-re) számítsuk ki:

$$\theta_{12}\hat{} = -\beta\hat{} \cdot \theta_{22}\hat{}, \text{ az}$$

$$1/\theta_{22}\hat{} = s_{22} - w_{12}^T\beta\hat{} \text{ egyenlet alapján}$$

#blockwise coordinate descent algorithm

eLasso: grafikus lasso bináris változókra

C. D. van Borkulo et al. (2014). A new method for constructing networks from binary data, *Scientific Reports* **4**, Article number: 5918, doi:10.1038/srep05918

$$\hat{\Theta}_j^{\rho} = \arg\min_{\Theta_j} \left\{ -x_{ij} \cdot \left(\tau_j + \sum_{k \in V_{\setminus j}} \beta_{jk} x_{ik} \right) + \log\left(1 + \exp\left\{ \tau_j + \sum_{k \in V_{\setminus j}} x_{ik} \beta_{jk} \right\} \right) + \rho \sum_{k \in V_{\setminus j}} \left| \beta_{jk} \right| \right\}$$

Két lényegi különbség:

- a változóként felírt regressziós modell logisztikus → Ising modell
- Θ mátrix particionálása eltér

Grafikus lasso implementáció - qgraph

Gráfstruktúra elemzésére szolgáló mértékek:

- Távolság (distance)
  - Mennyire közvetlenül befolyásol egy csomópont egy másikat?
- Centralitás (centrality)
  - Melyik csomópontnak van ‘központi szerepe’?
- Összekötöttség (connectivity)
  - Mennyire vannak összeköttetésben a csomópontok?

qgraph - távolság

Távolság (distance)

- Mennyire közvetlenül befolyásol egy csomópont egy másikat?

- Súlyozott éleknél: a távolság a súlyokkal fordítottan arányos
  - Ha egy élnek nagy súlya van: 'közel' van egymáshoz a két csomópont
- Legrövidebb út két csomópont között
- Távolság számítása:
  - Súlyozott esetben az úton lévő élsúlyok összege
  - Súlyozatlan esetben az utat alkotó élek száma

qgraph - centralitás

- Fokszám / erősség (strength)
  - A ‘legerősebb’ csomópont közvetlenül prediktálja a legtöbb csomópontot
- Közelség (closeness)
  - Az a csomópont, amely a legjobb predikciós értékkel bír más (nem közvetlen) csomópontokra
  - ‘Közel’ van a legtöbb csomóponthoz
- Összekötöttség – közbülsőség (betweenness)
  - Összeköttetésben áll a fontosabb csomópontokkal

qgraph -glasso

- `library("qgraph")`
- `library("psych")`
- `data(bfi)`
- `bfi ← bfi[,1:25]`
- `cor_bfi ← cor_auto(bfi)`
- `graph_cor ← qgraph(cor_bfi, layout = "spring")`

Parciális korrelációk:

- `graph_pcor ← qgraph(cor_bfi, graph = "concentration", layout = "spring")`

qgraph -glasso

Grafikus lasso (glasso)

- `graph_glas <- qgraph(cor_bfi, graph = "glasso", sampleSize = nrow(bfi))`

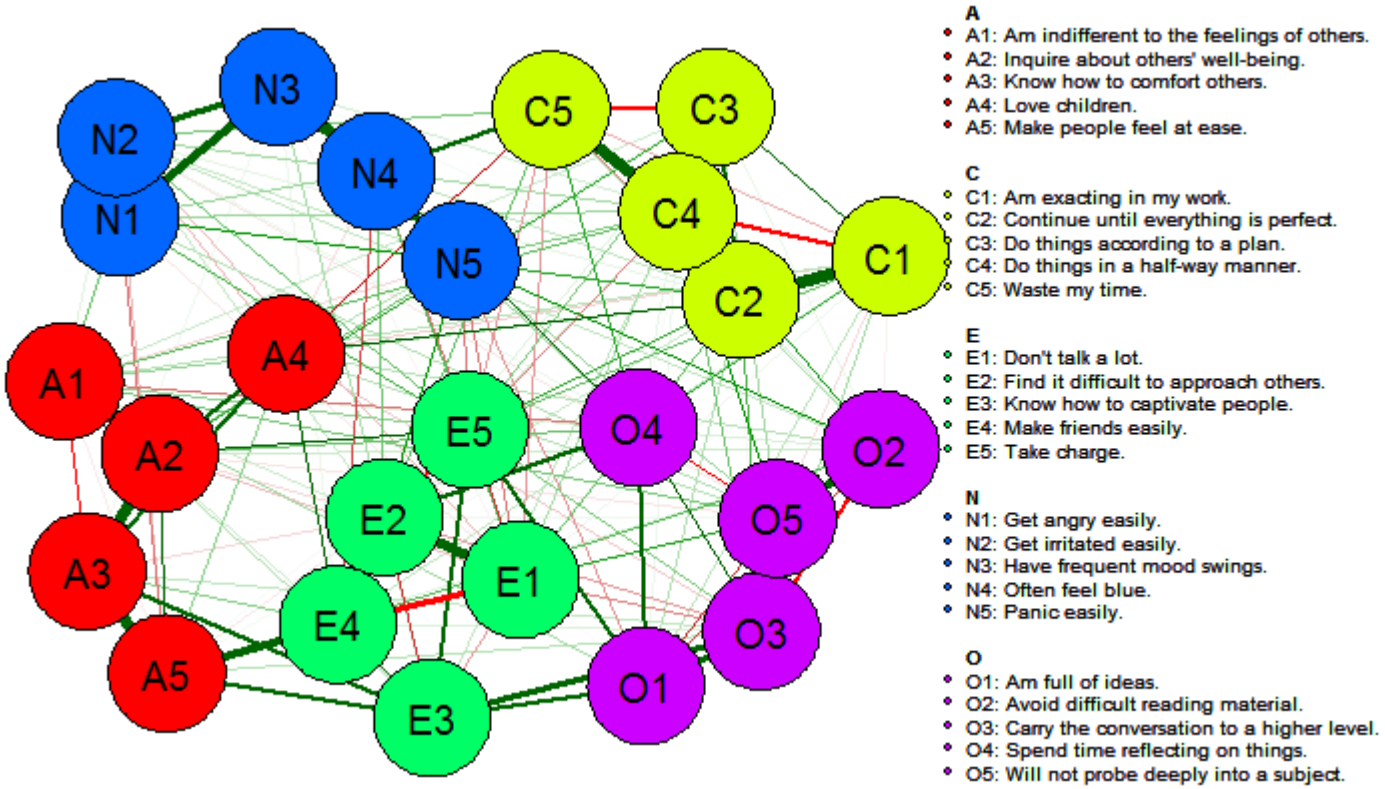
- `Names <- scan ("http://sachaepskamp.com/files/BFIitems.txt", what = "character", sep = "\n")`
- `Groups <- rep(c('A','C','E','N','O'),each=5)`

Grafikus lasso (glasso) + jelmagyarázat

- `qgraph(cor_bfi, graph = "glasso", sampleSize = nrow(bfi),layout = "spring", nodeNames = Names,legend.cex =0.6, groups = Groups)`

BFI - glasso

- BFI data set



qgraph –glasso – centralitás értékek

- Closeness
- Betweenness
- Strength

```
centralityPlot(graph_glas, labels=Names)
centralityPlot(list(cor = graph_cor, pcor =
  graph_pcor, glasso = graph_glas))
```

BFI - glasso

- Csomópont centralitások