Analysis of a biomedical knowledge graph

E. Mosca – 276661

C. El Arrag - 276931

Initial Analysis of the Graph

Grasping what the data looks like

Total number of entities(nodes) for each kind

```
kind
Gene                  20945
Biological Process    11381
Side Effect            5734
Molecular Function     2884
Pathway                1822
Compound               1552
Cellular Component     1391
Symptom                 438
Anatomy                 402
Pharmacologic Class     345
Disease                 137
Name: count, dtype: int64
```

Average in-degree and out-degree for each type of node

	In-degree	Out-degree
Side Effect	24.231601	0.000000
Biological Process	49.161234	0.000000
Pharmacologic Class	0.000000	2.982609
Symptom	7.664384	0.000000
Gene	54.452232	61.507424
Compound	4.842139	127.538015
Pathway	46.307355	0.000000
Molecular Function	33.710818	0.000000
Cellular Component	52.887132	0.000000
Disease	12.321168	256.948905
Anatomy	8.960199	1462.261194

Finding in and out degrees for diseases

```
hypertension : in-degree = 73
hematologic cancer : in-degree = 53
breast cancer : in-degree = 44
asthma : in-degree = 41
coronary artery disease : in-degree = 40
```

```
breast cancer : out-degree = 1079
hematologic cancer : out-degree = 1033
IgA glomerulonephritis : out-degree = 948
melanoma : out-degree = 917
rheumatoid arthritis : out-degree = 842
```

Finding most connected diseases

```
breast cancer : total degree = 1123
hematologic cancer : total degree = 1086
IgA glomerulonephritis : total degree = 948
melanoma : total degree = 930
rheumatoid arthritis : total degree = 880
```

*Highest total-degs (in-deg + out-deg)

Total Number of Unique Edgetype

```
metaedge
GpBP     559504
AeG      526407
Gr>G     265672
GiG      147164
CcSE     138944
AdG      102240
AuG       97848
GpMF      97222
GpPW      84372
GpCC      73566
GcG       61690
CdG       21102
CuG       18756
DaG       12623
CbG       11571
DuG        7731
DdG        7623
CrC        6486
DlA        3602
DpS        3357
PCiC       1029
CtD         755
DrD         543
CpD         390
Name: count, dtype: int64
```

Defining Similarity Measures (Diseases)

Many things may lead to two nodes being similar:

- Basic measure: shared neighbors/successors/predecessors

For diseases:

- Shared Symptoms between diseases
- Shared genes between diseases
- Shared compounds palliating(CpD) or treating(CtD) a disease
- Connections between diseases (the DrD edgetype says that a disease resembles another one)

Example: similarities between Type 1 and Type 2 diabetes melitus

		Edge Connecting Diseases	Shared Compounds Cpd	Shared Compounds Ctd	Shared Genes	Shared Neighbors	Shared Predecessors	Shared Similar Diseases	Shared Symptoms
type 1 diabetes mellitus	type 2 diabetes mellitus	True	-	0.045 (1 / 22)	0.084 (33 / 392)	0.133 (67 / 505)	0.061 (2 / 33)	0.1 (1 / 10)	0.138 (65 / 472)

Defining Similarity Measures (Symptoms)

Characteristics of similar symptoms:

- Leading to similar diseases (shared diseases)
- Shared anatomy through diseases (shared affected body parts through diseases)

Example: similarities between eye pain and blindness

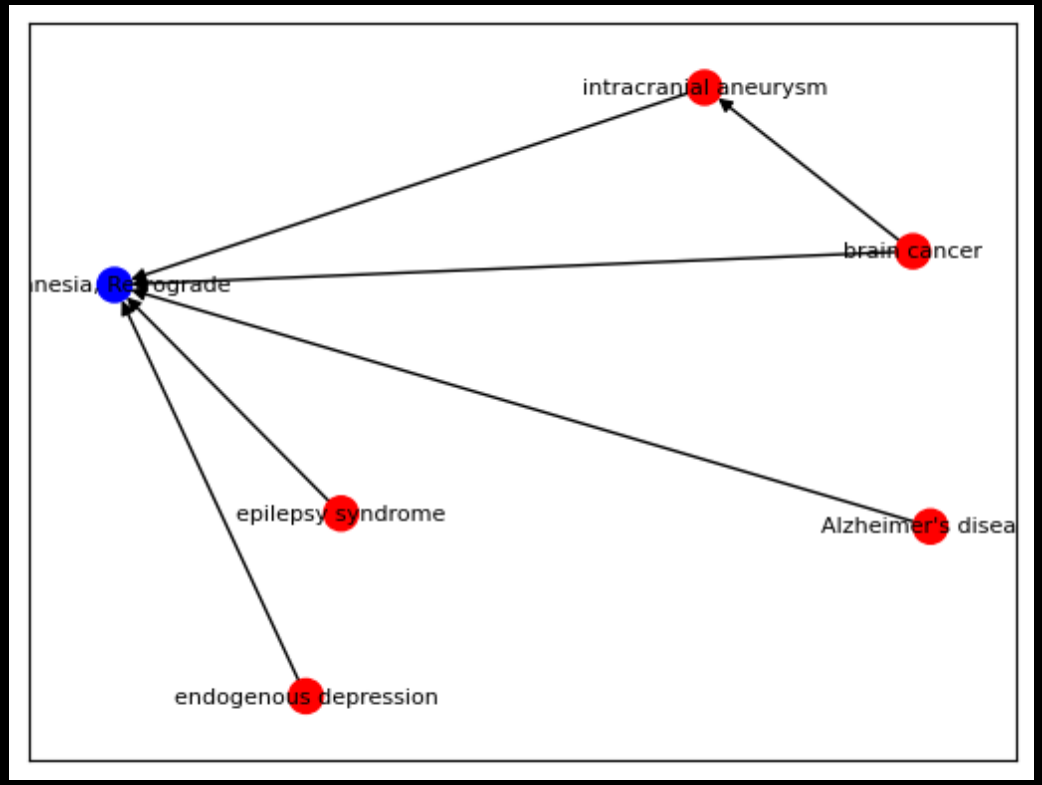
		Shared Anatomy	Shared Diseases
Eye Pain	Blindness	0.705 (249 / 353)	0.35 (7 / 20)

Patient Treatment: Retrograde Amnesia

«The inability to access memories or information from before an injury or disease occurred»

For a given symptom, there might be many diseases pointing to it, in our graph this means that a disease presents a symptom, and we want to evaluate each one in order to make an educated guess or provide more info to the decision maker. Some measures could be:

- Number of connected symptoms: a disease with more possible symptoms lowers the likelihood.
- Number of connected genes: similarly, more genes, less likelihood.
- Betwenness Centrality: how involved is the disease in connecting nodes? i.e. How many (shortest paths) may lead us to the symptom through the disease?



	connected_symptoms	connected_genes	betweenness_centrality
intracranial aneurysm	50.0	23.0	3.590837e-07
Alzheimer's disease	44.0	685.0	2.257420e-04
epilepsy syndrome	64.0	399.0	1.319269e-03
brain cancer	88.0	111.0	7.810917e-06
endogenous depression	53.0	33.0	1.711393e-04

Patient Treatment: Suggesting Compounds

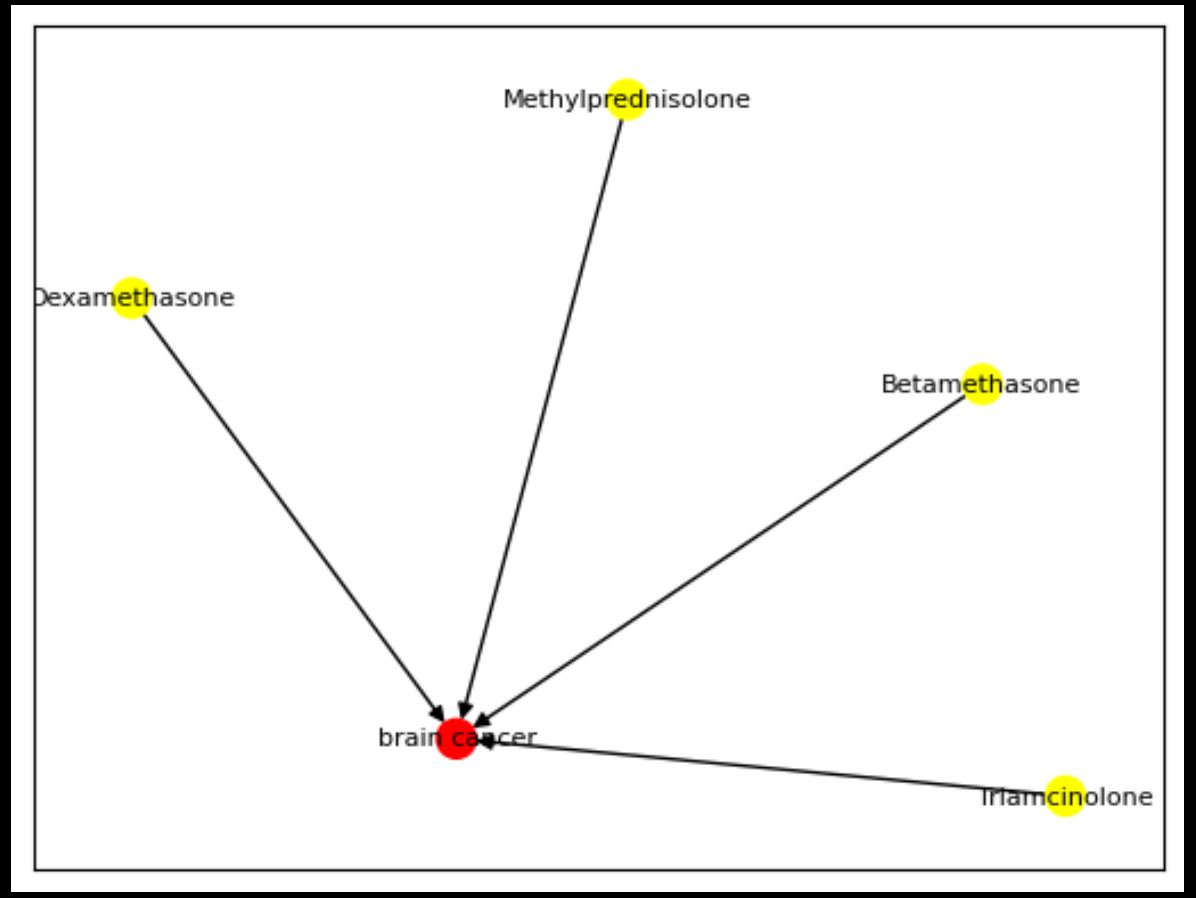
Among the possible diseases, we decide to further analyse the brain cancer case.

Therefore we move on to visualize possible compounds which could alleviate significant symptoms of the disease(CpD), and look for compounds which might help the patient treat the disease as a whole(CtD)

Regarding disease treatment, we find the following compounds:

Compound ID	Number of Possible side-effects
Compound::DB00853	241
Compound::DB00262	177
Compound::DB00773	153
Compound::DB01168	137
Compound::DB01005	130
Compound::DB01206	44
Compound::DB00958	7

It follows that we want to suggest the set of compounds with the lowest number of side effects

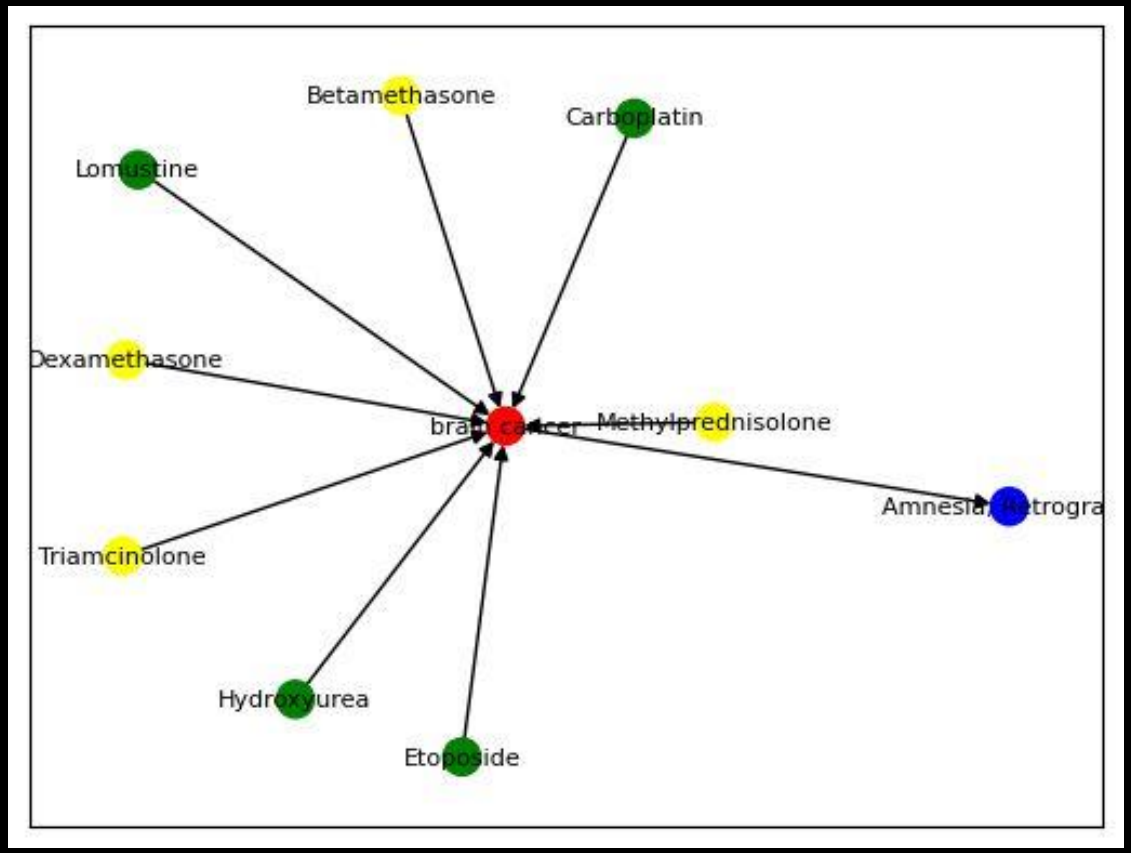


Patient Treatment: Selected Compounds

Among all disease-treating compounds, we look for the set with the lowest number of unique side effects, with one compound for each pharmacologic class.

From the previous set of compounds we found the following, reducing the number of possible side effects to 253.

	Selected Compound
Platinum-containing Compounds	Carboplatin
Alkylating Activity	Lomustine
Urea	Hydroxyurea
Topoisomerase Inhibitors	Etoposide




Thanks for your time!
E. Mosca
C. El Arrag