# The Hitchhiker's Guide to Graph Exchange Formats

Prof. Matthew Roughan
matthew.roughan@adelaide.edu.au
http://www.maths.adelaide.edu.au/matthew.roughan/
Work with Jono Tuke

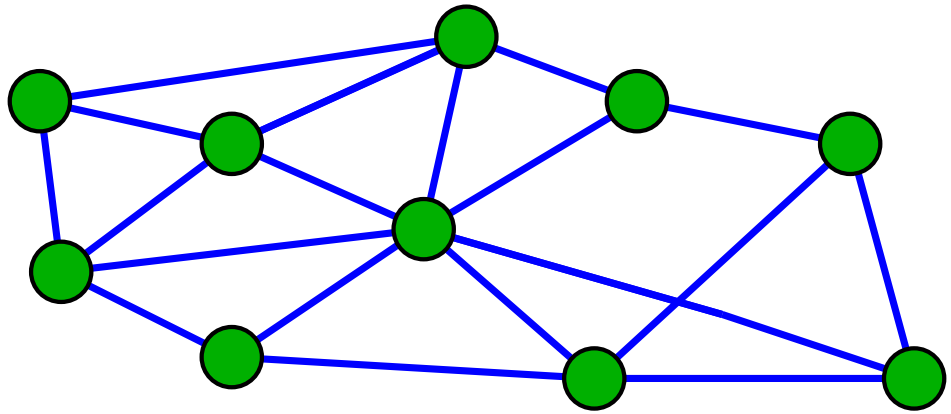
UoA

June 4, 2015


THE UNIVERSITY of ADELAIDE

# Graphs

- Graph: $G(N, E)$
  - $N$ = set of nodes (vertices)
  - $E$ = set of edges (links)

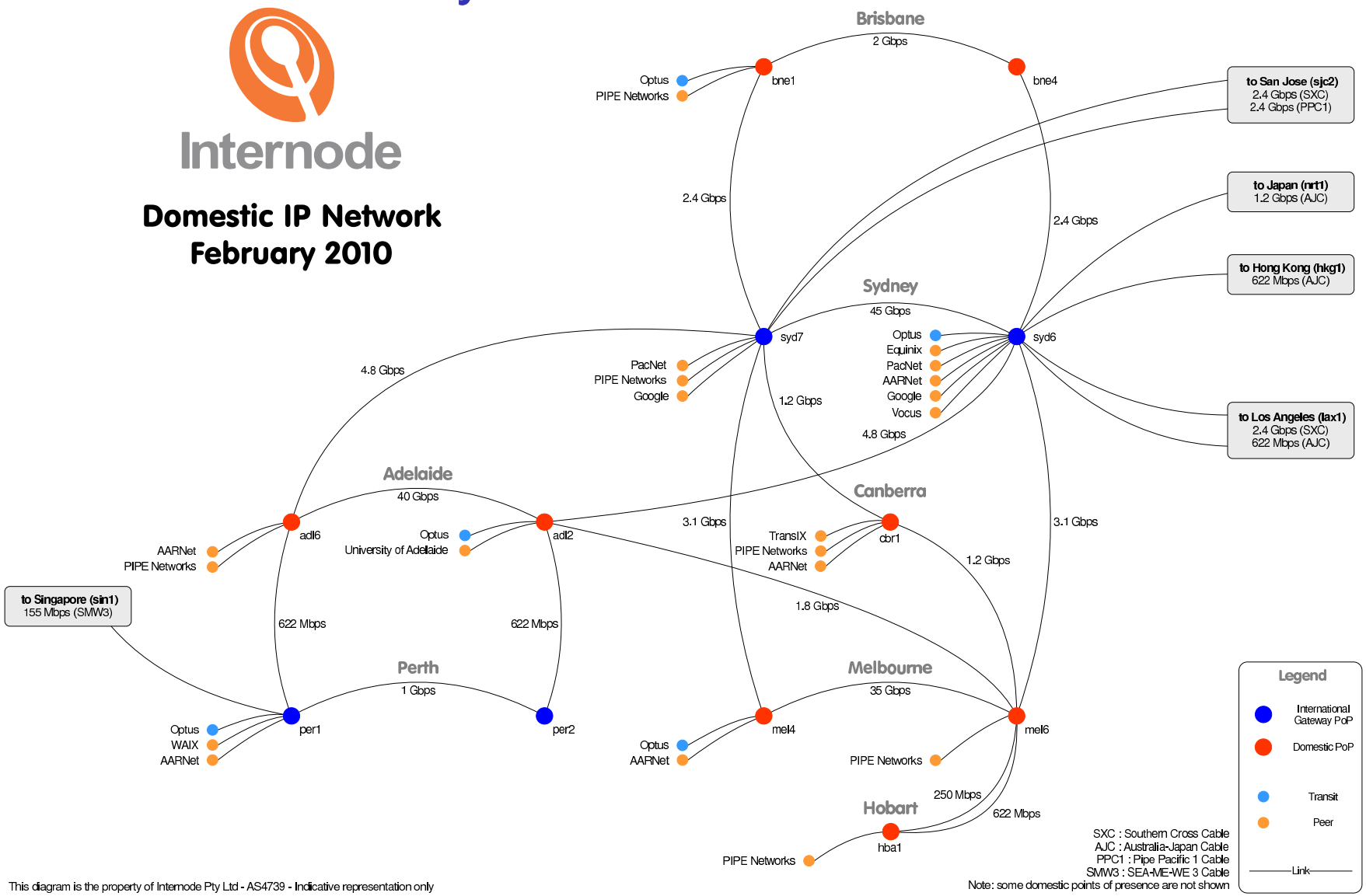


- Often we have additional information, e.g.,
  - link distance
  - node type
  - graph name

# Why?

- To represent data where “connections” are 1st class objects in their own right
  - storing the data in the right format improves access, processing, ...
  - it’s natural, elegant, efficient, ...
- Many, many datasets

# ISPs: Internode: layer 3

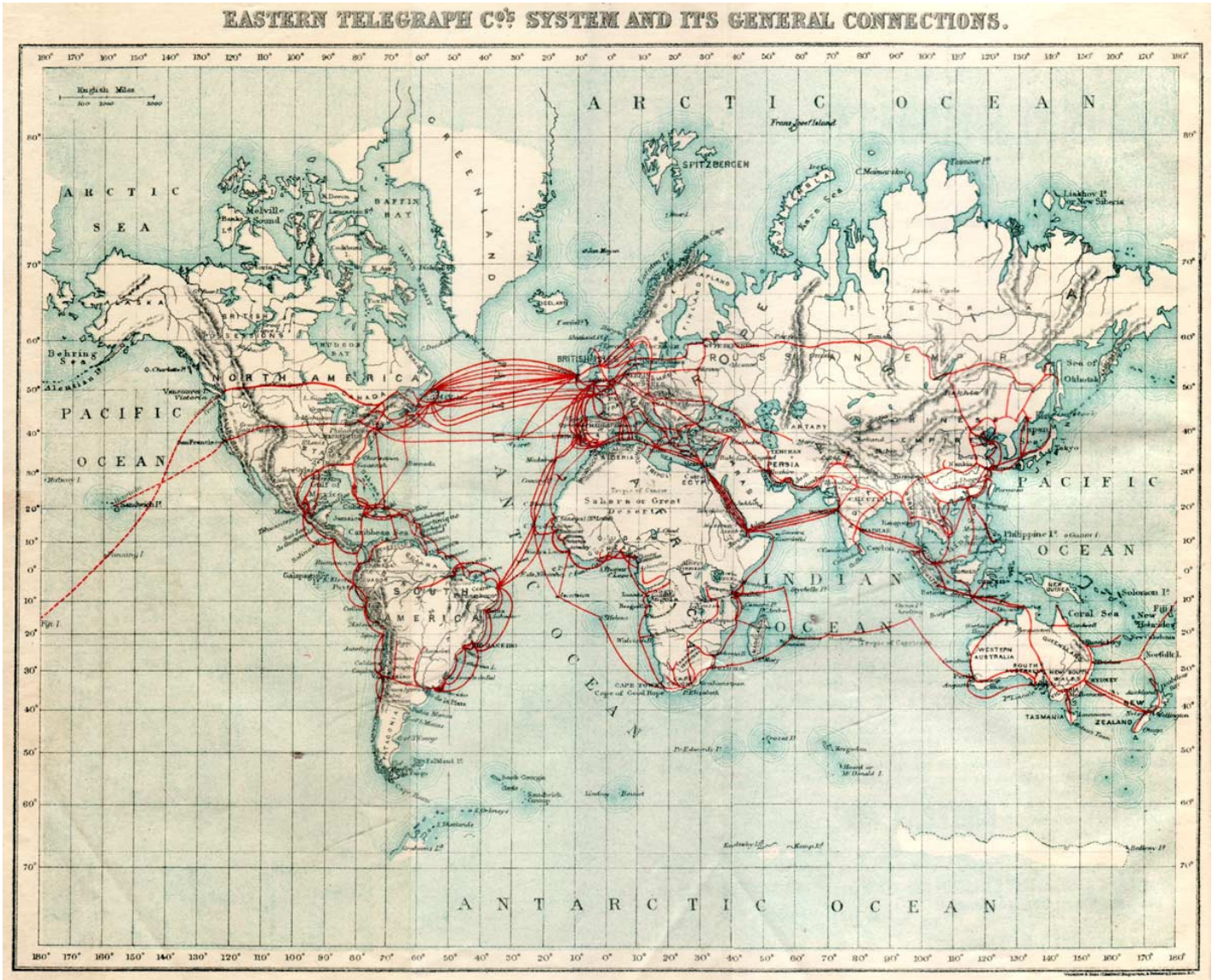



http://www.internode.on.net/pdf/network/internode-domestic-ip-network.pdf

# ISPs: Level 3 (NA)

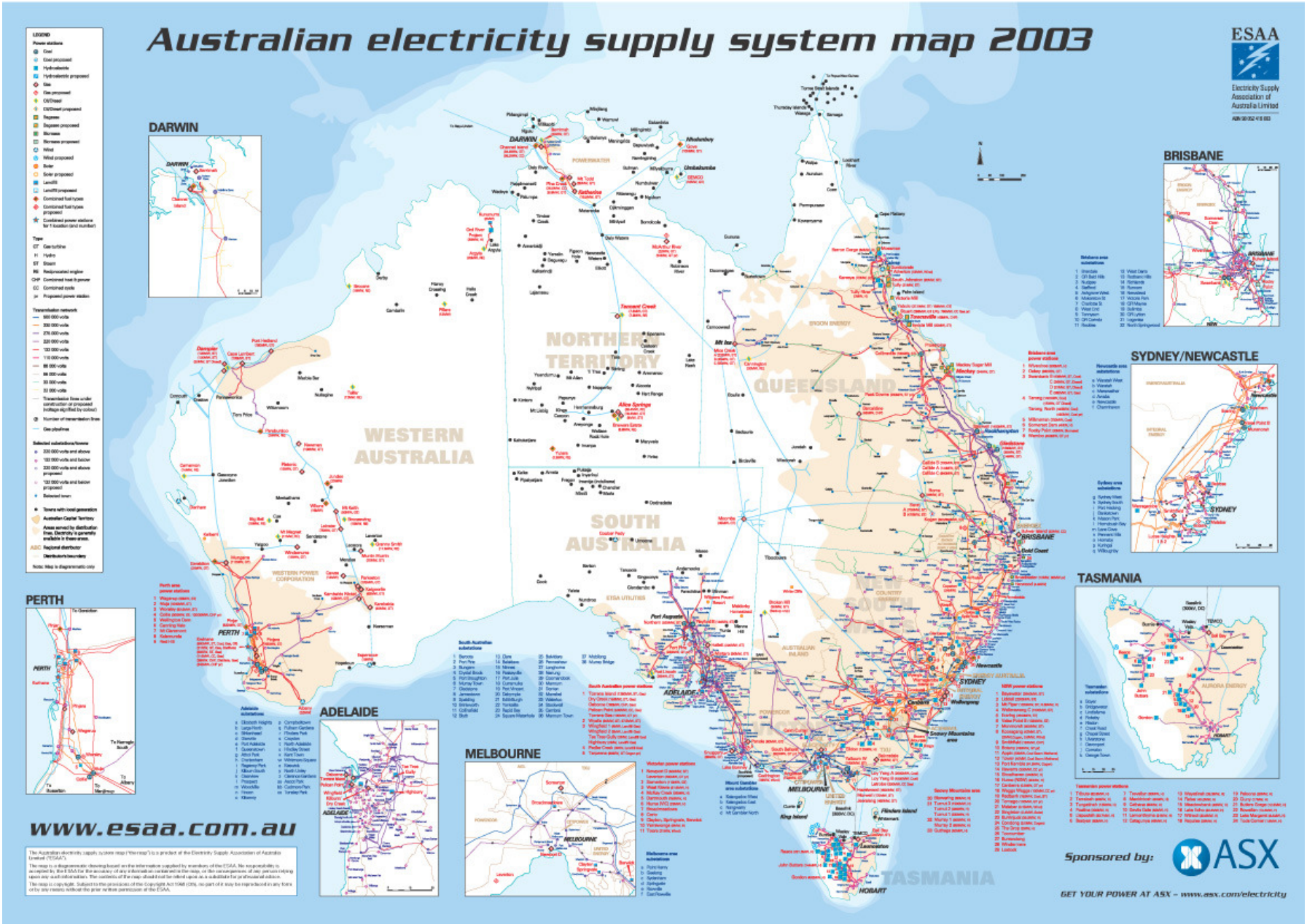

http://www.fiberco.org/images/Level3-Metro-Fiber-Map4.jpg

# Telegraph submarine cables

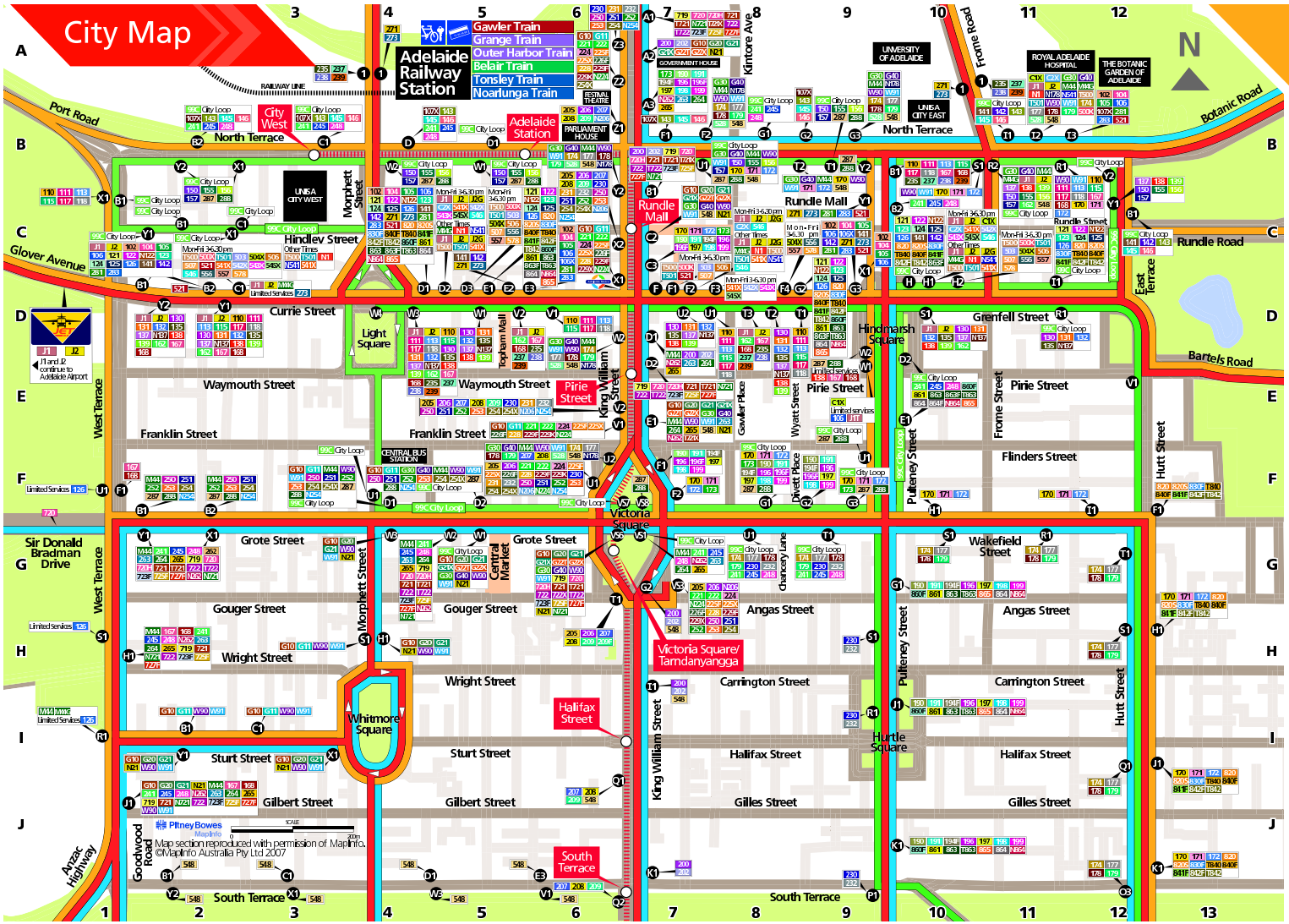



http://en.wikipedia.org/wiki/File:1901_Eastern_Telegraph_cables.png

# Electricity grid



# Bus network (Adelaide CBD)




## FREE City Services

**Free bus and tram services around the City, 7 days a week**

Travel around Adelaide City centre by either the FREE 99C City Loop bus or the FREE Terrace to Terrace tram services, both linking you to many major attractions and transfers to other Adelaide Metro services.

All 99C and FREE Terrace tram service **access** for everyone

### Adelaide FREE 99C City Loop buses run in both d

**Monday to Friday:** 8 am to 6.15 pm - every 15 minutes
**Friday evening:** 6.15 pm to 9.15 pm - every 30 minutes
**Saturday:** 8.15 am to 5.45 pm - every 30 minutes
**Sunday and public holidays:** 10.15 am to 5.45 pm - every 30

### FREE Terrace to Terrace Tram Service

To travel on the FREE Terrace to Terrace tram, board either t
South Terrace tram services departing:

**Monday to Friday:** 8 am to 6 pm - every 7.5 minutes on aver
**Saturday, Sunday and public holidays:** 9 am to 6 pm - every approximately
**Other times to Midnight:** Every 20 minutes approximately
**For travel between South Terrace and Glenelg you'll need a**

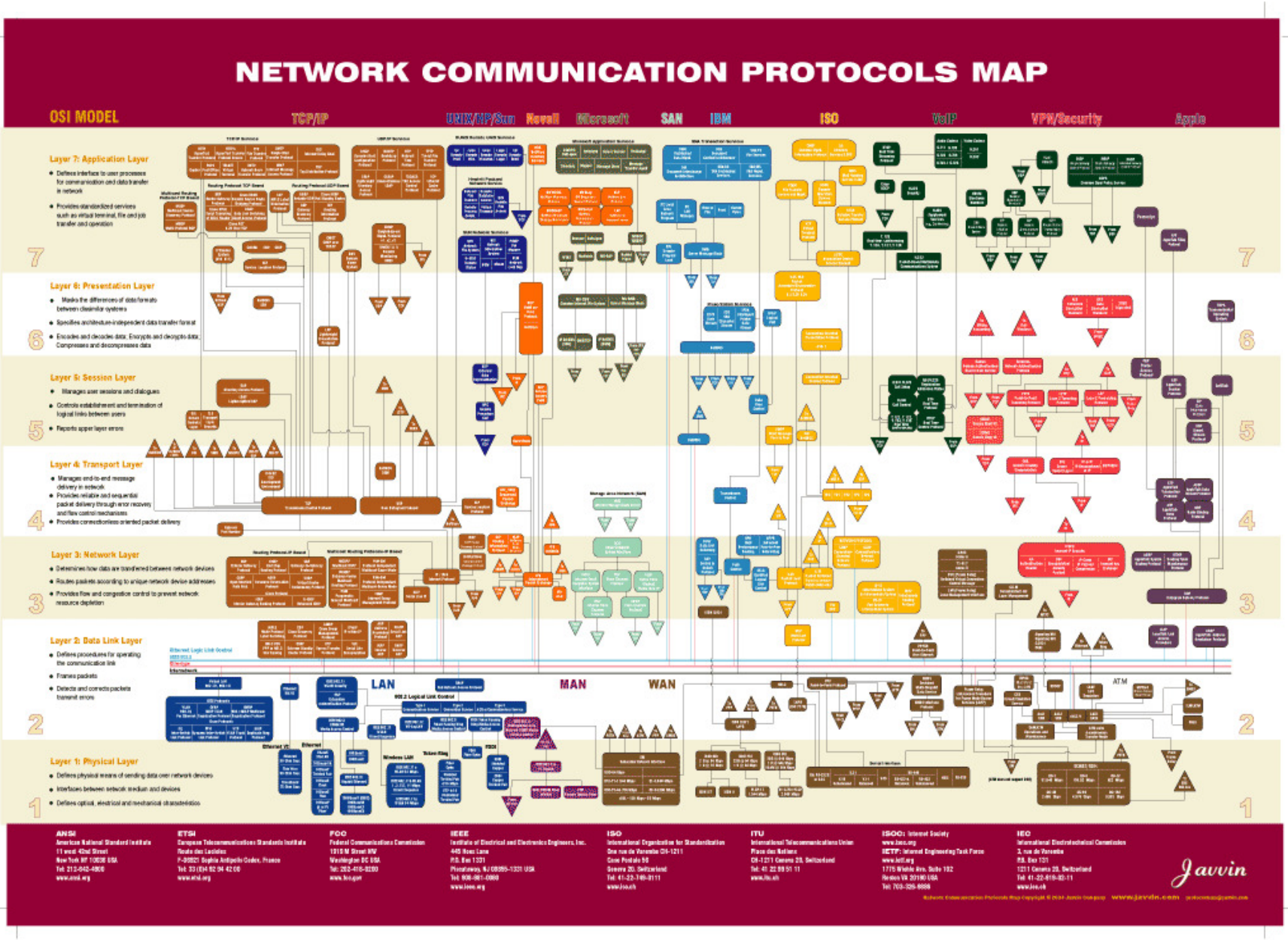
### City Map Legend

| | | | |
|---|---|---|---|
| | Adelaide Metro InfoCentre | | Bike Lockers a |
| X1 | Bus stop | | Ticket Sales |
| | Railway Line | | JetBus direct li Airport, 7 days |
| 99C | 99C City Loop Adelaide FREE bus | | Accessible bus Adelaide FREE FREE trams and of all Adelaide services are full |
| | Tramline and stops | | |
| 170 171 | Routes available from that stop | | |

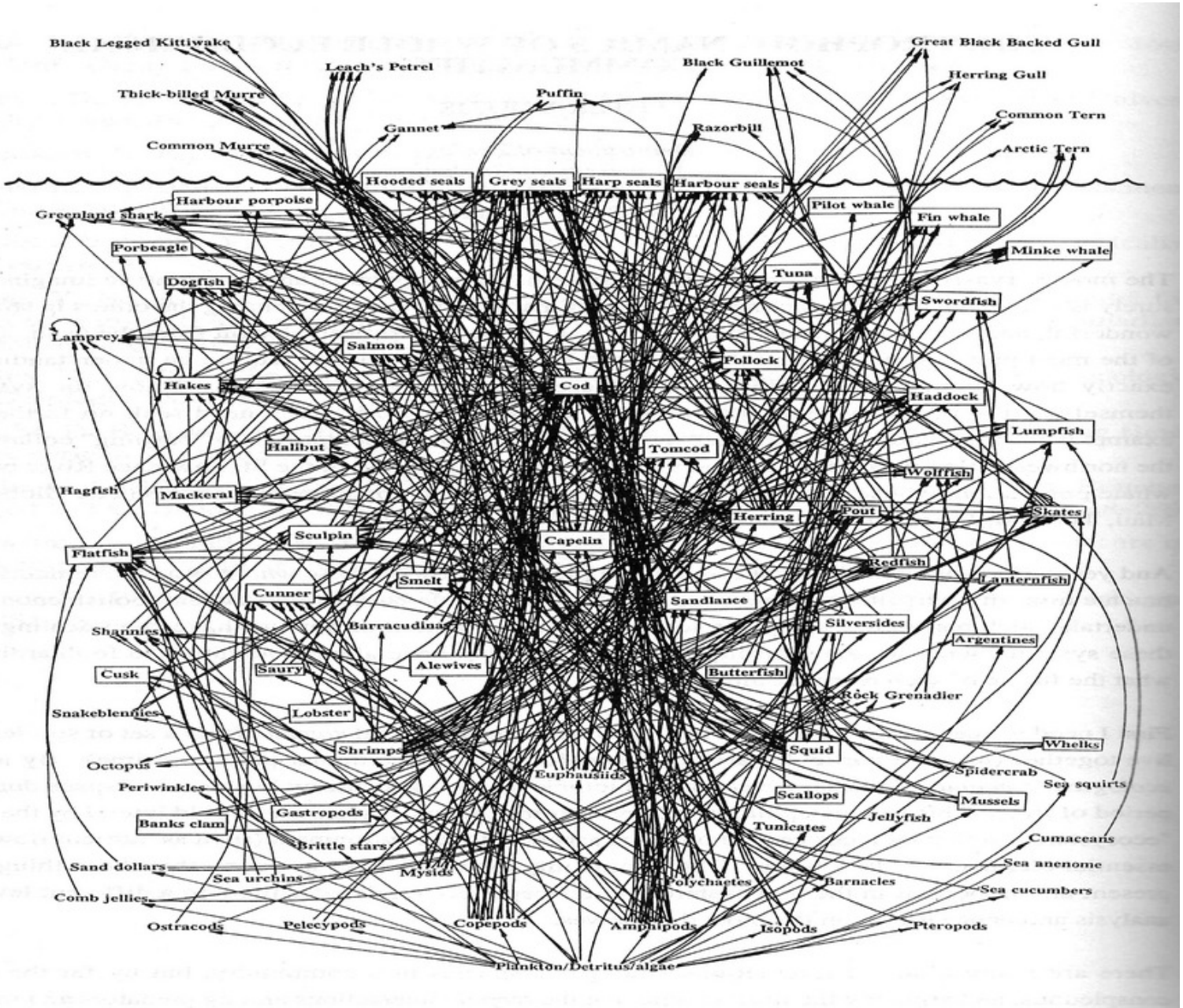
# French Rail


http://www.alleuroperail.com/europe-map-railways.htm

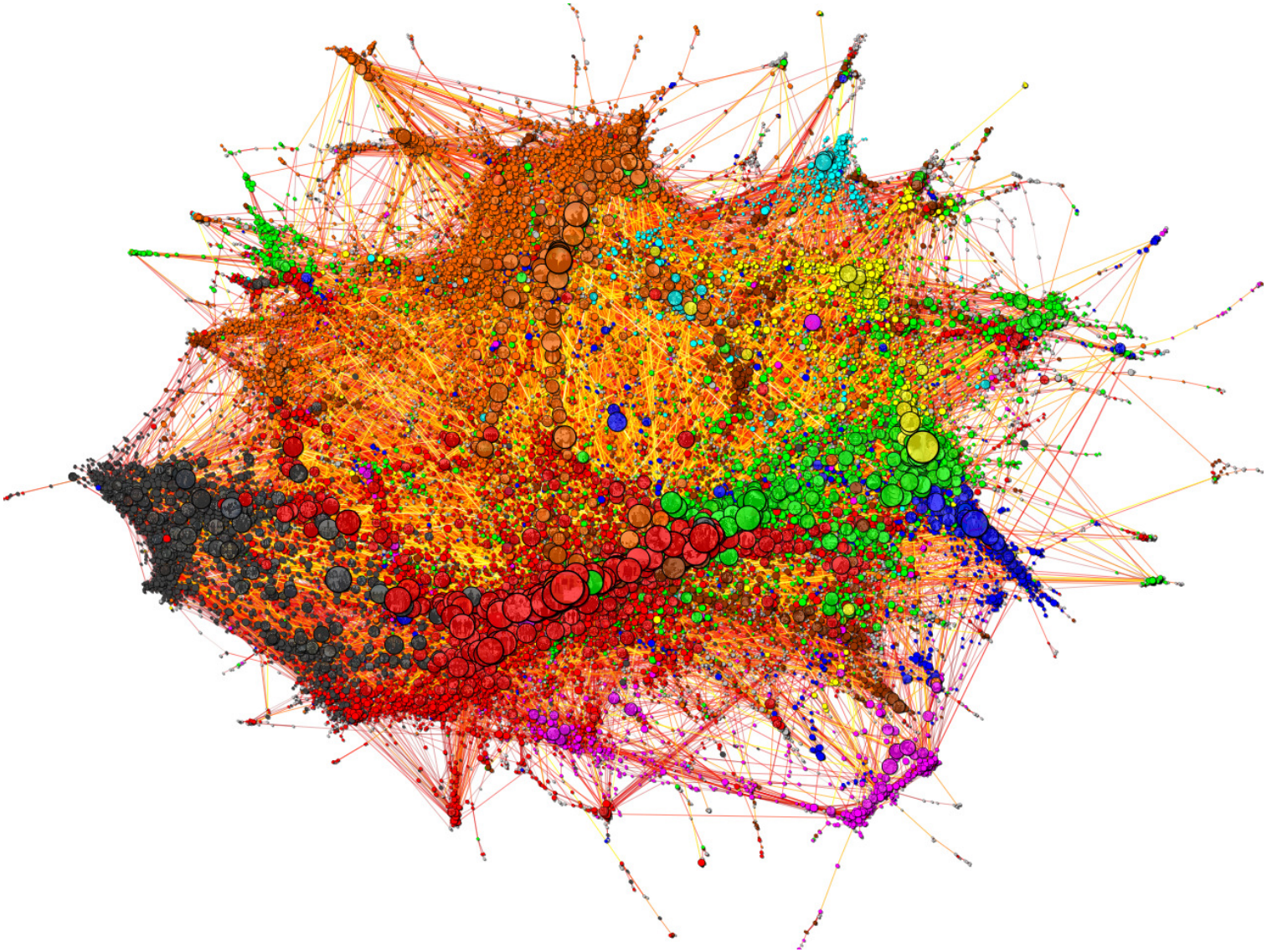
# Protocol relationships



# Food web


Black Legged Kittiwake
Leach's Petrel
Thick-billed Murre
Puffin
Gannet
Razorbill
Black Guillemot
Great Black Backed Gull
Herring Gull
Common Tern
Arctic Tern
Common Murre
Hooded seals
Grey seals
Harp seals
Harbour seals
Harbour porpoise
Greenland shark
Pilot whale
Fin whale
Porbeagle
Minke whale
Tuna
Dogfish
Swordfish
Lamprey
Salmon
Pollock
Hakes
Cod
Haddock
Lumpfish
Halibut
Tomcod
Wolffish
Hagfish
Mackeral
Herring
Pout
Skates
Sculpin
Capelin
Flatfish
Redfish
Lanternfish
Smelt
Cunner
Sandlance
Barracudina
Silversides
Shannies
Argentines
Alewives
Saury
Butterfish
Cusk
Rock Grenadier
Snakeblennies
Lobster
Whelks
Shrimps
Squid
Octopus
Euphausiids
Spidercrab
Periwinkles
Sea squirts
Scallops
Mussels
Gastropods
Jellyfish
Banks clam
Tunicates
Brittle stars
Cumaceans
Sand dollars
Sea urchins
Mysids
Sea anemone
Polychaetes
Barnacles
Comb jellies
Sea cucumbers
Ostracods
Pelecypods
Copepods
Amphipods
Isopods
Pteropods
Plankton/Detritus/algae

# Network of Musicians (last.fm)



http://sixdegrees.hu/last.fm/

# Exchange

- Open research requires exchange of data
  - replications of, and comparisons with results
  - comparisons between datasets
- Working on data in closed formats is bad
  - vendor lock-in
  - not free
  - not portable (or sometimes even backward compatible)
  - often black boxes
- Exchange formats are designed to facilitate open exchange of data

# Portability

Main requirement is portability

- Portability between software
  - graph entry
  - graph analysis
  - graph visualisation
  - ...
- Portability between architecture
  - OS (Mac, Linux, Windows, FreeBSD, ...)
  - Hardware (big-endian v little-endian)

Requirements

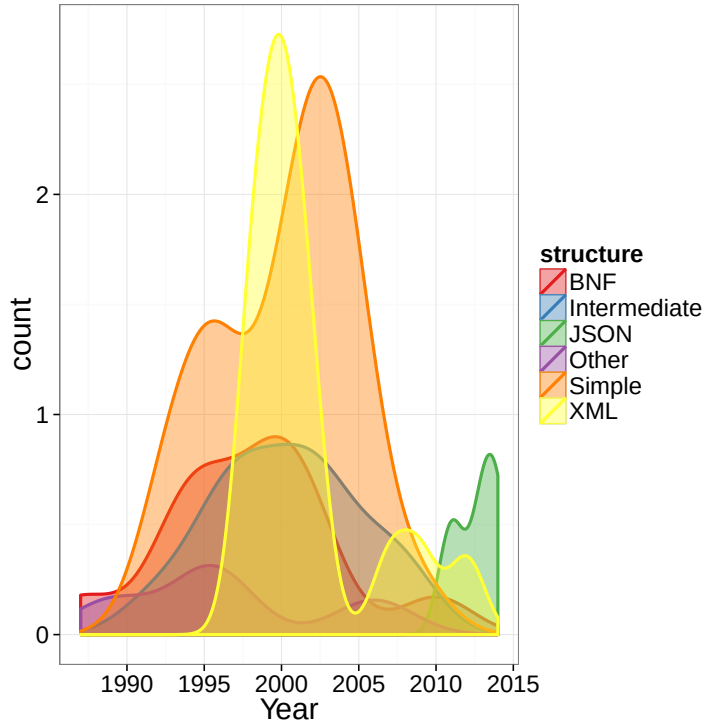
- Open format
- Documented

# Graph exchange file formats

| | | | |
|---|---|---|---|
| 1 | bintsv4 | | bintsv4 (GraphLab) |
| 2 | BioGRID TAB | ✓ | BioGRID TAB 2.0 Format |
| 3 | BLAG, GDToolkit | | Batch layout generator (GDToolkit) |
| 4 | BVGraph | ✓ | Boldi-Vigna graph compression |
| 5 | Chaco | ✓ | Chaco graph format |
| 6 | Cluto | | Cluto/Metis/Graclus format |
| 7 | DGS | ✓ | Dynamic GraphStream Format |
| 8 | DGML | | Directed Graph Markup Language |
| 9 | DIMACS | | DIMACS graph format |
| 10 | Dot | | GraphVis Dot Language |
| 11 | DotML | | Dot Markup Language |
| 12 | DyNetML | | DyNetML XML |
| 13 | GAMFF | | A Graph and Matrix Format |
| 14 | GDF | | Guess Data Format |
| 15 | GDL | | Graph Description Language |
| 16 | GEDCOM | | Geneaological data |
| 17 | GEXF | ✓ | Graph Exchange XML Format |
| 18 | GML | ✓ | Graph Modelling Language |
| 19 | Graph6 | ✓ | Graph6 |
| 20 | Graph::Easy | ✓ | Perl Graph::Easy format |
| 21 | GraphEd | | GraphEd simple format |
| 22 | GraphJSON | | Graph JSON |
| 23 | GraphML | ✓ | Graph Markup Language |
| 24 | GraphSON | | TinkerPop's JSON-based Graph format |
| 25 | GraphXML | ✓ | XML-Based Graph Description Language |
| 26 | GraX | | GraX |
| 27 | GRXL | | XML Specification for Grrr Program |
| 28 | GT-ITM | | Georgia Tech Internetwork Topology Models |
| 29 | GXL | ✓ | Graph eXchange Language |
| 30 | Harwell-Boeing | | Harwell-Boeing sparse (TGFaceny) matri |
| 31 | Inet | | Inet Topology Generator file |
| 32 | ITDK | ✓ | CAIDA Internet Topology Data Kit |
| 33 | JSON Graph | | json-graph-specification |
| 34 | LEDA | | LEDA format |
| 35 | LGF | | LEMON Graph Format |

# Graph exchange file formats

- Over 100 considered
- 76 analysed
  - aiming at exchange formats
    - ★ not all designed for exchange, but some became *de facto* exchange formats
  - sought minimal level of documentations
    - ★ not all exchange formats are documented (still)

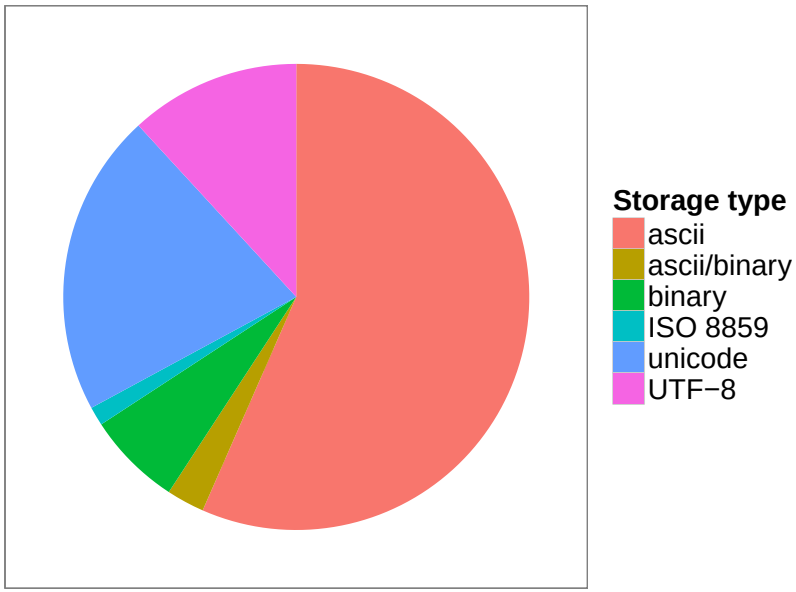


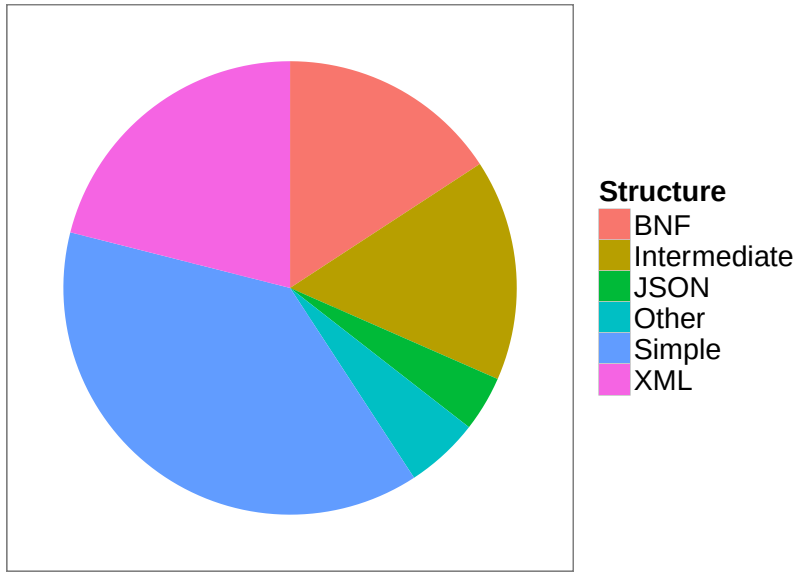
# Descriptors

Many possibilities (considered 26)

- File type
  - encoding
  - representation
  - meta-data
  - compression
  - ...
- Graph types
  - directed, multi-, hyper-, ...
- Attributes
  - what attributes can be stored with graph
  - extensibility
- General
  - extensibility
  -

# Encoding Types



# Graph Representations

- List of links: explicitly give
  - $N$: e.g., $N = \{1, 2, 3\}$
  - $E$: e.g., $E = \{(1, 2), (1, 3)\}$
- Adjacency matrix: define connectivity through a $(0, 1)$ matrix $A$ defined by

$$A_{ij} = \begin{cases} 1, & \text{if } (i, j) \in E \\ 0, & \text{otherwise} \end{cases}$$

  e.g.,

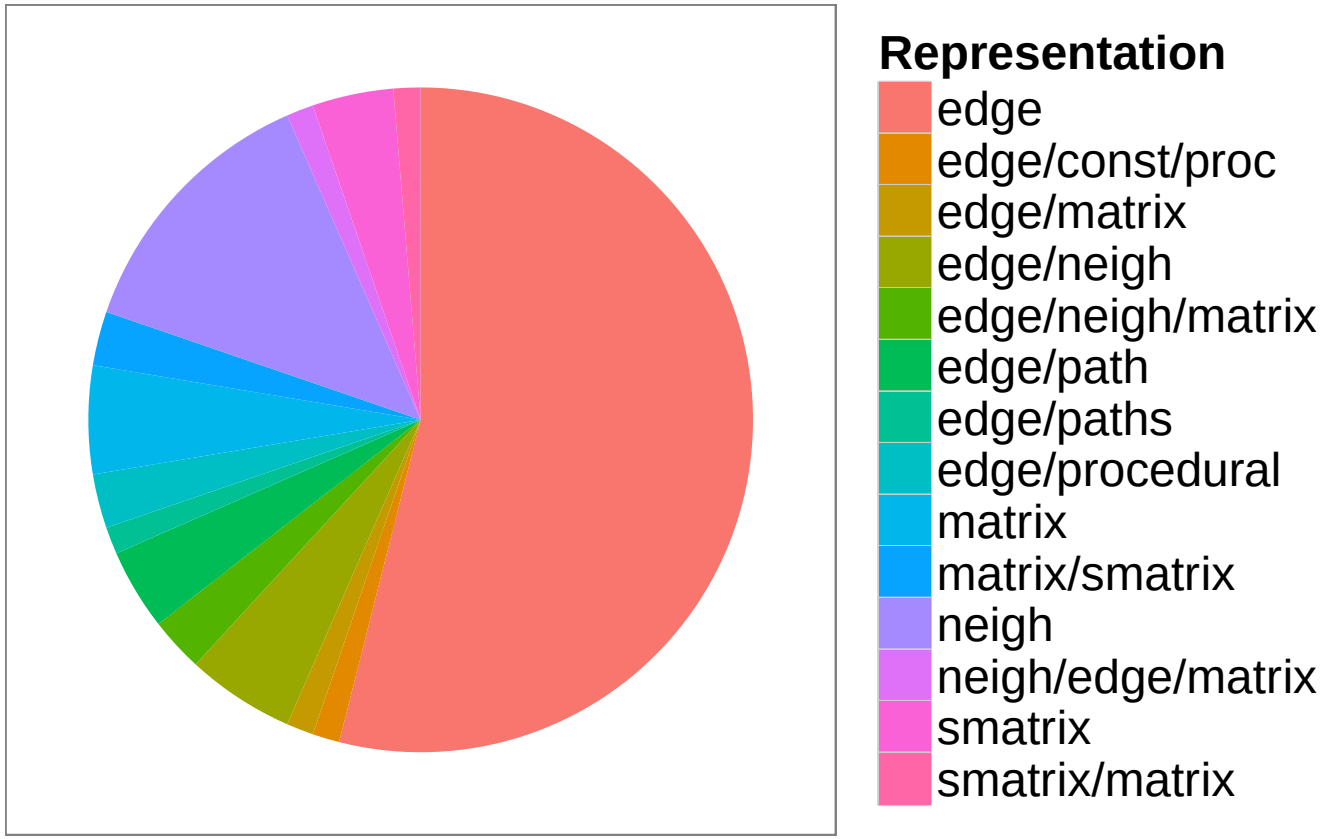
$$A = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}$$

- List of neighbour lists:
  - for each node: list its neighbours
  - e.g.

| | |
|---|---|
| 1 | 2, 3 |
| 2 | |
| 3 | |

- Paths
- Constructive and/or procedural

# Graph Representations



# Descriptors: Graph Types and Generalisations

- directed/undirected
- acyclic/trees
- multigraph or pseudograph: has multiple parallel links between two nodes
  - e.g. its easy to have two links between two routers
  - also allows self-loop
- hypergraph: links connect more than two nodes
  - e.g., where you have a connective medium (rather than a wire), for instance in a wireless network.
- hierarchy
  - nodes have subgraphs
- meta-graph

# Descriptors: Attributes

| | Static | Dynamic |
|---|---|---|
| Node | name, location, type, ... | up/down, ... |
| Edge | distance, capacity, ... | up/down, utilisation, ... |

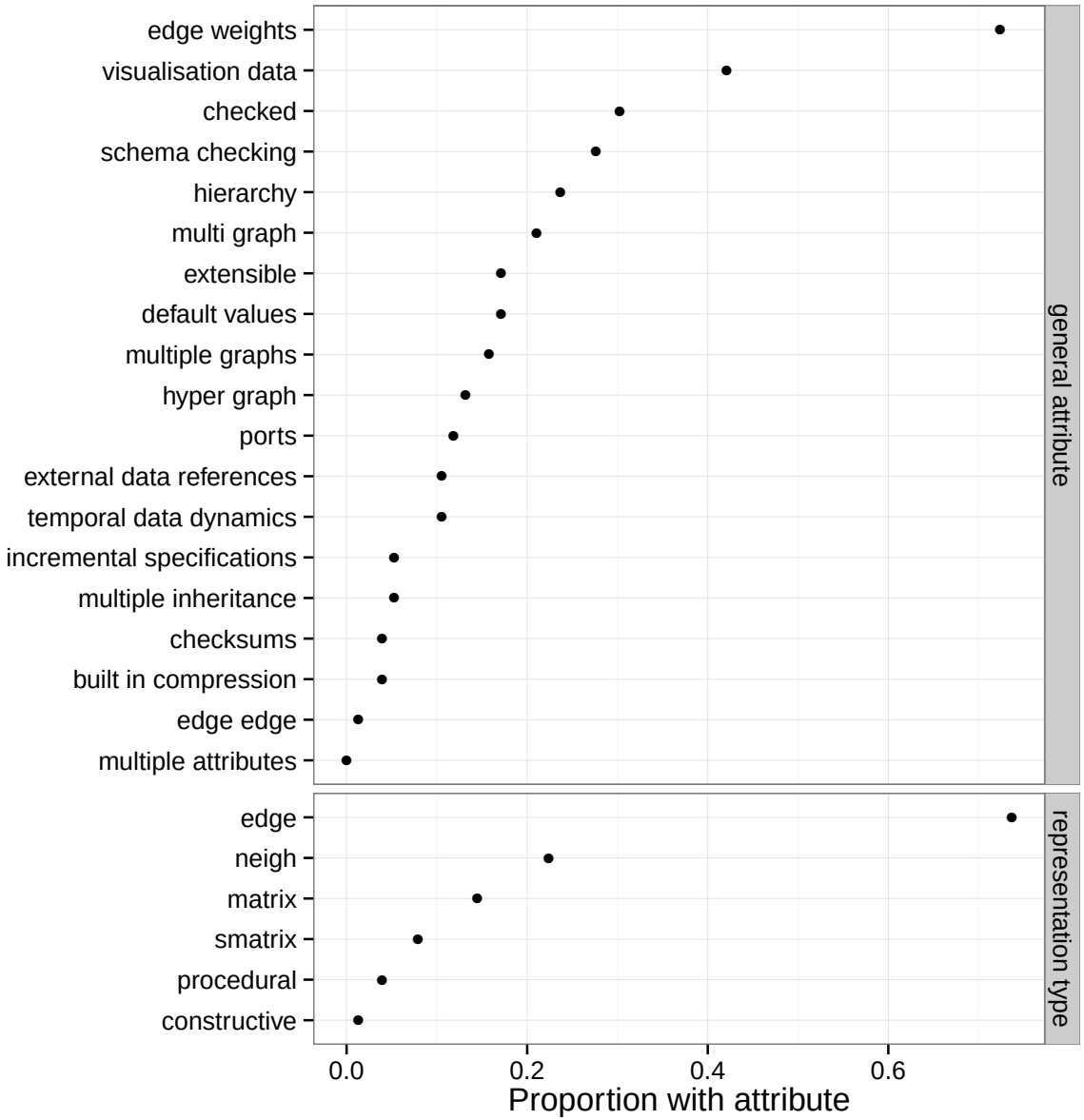
# Descriptors: Attributes

- Single number/weight (per edge)
- Multiple
  - fixed vs extensible
- Defaults
  - multiple inheritance
- Visualization
  - not really attributes of graph
  - used for drawing it
  - color, shape, layout, ...
- Ports
- Temporal dynamics

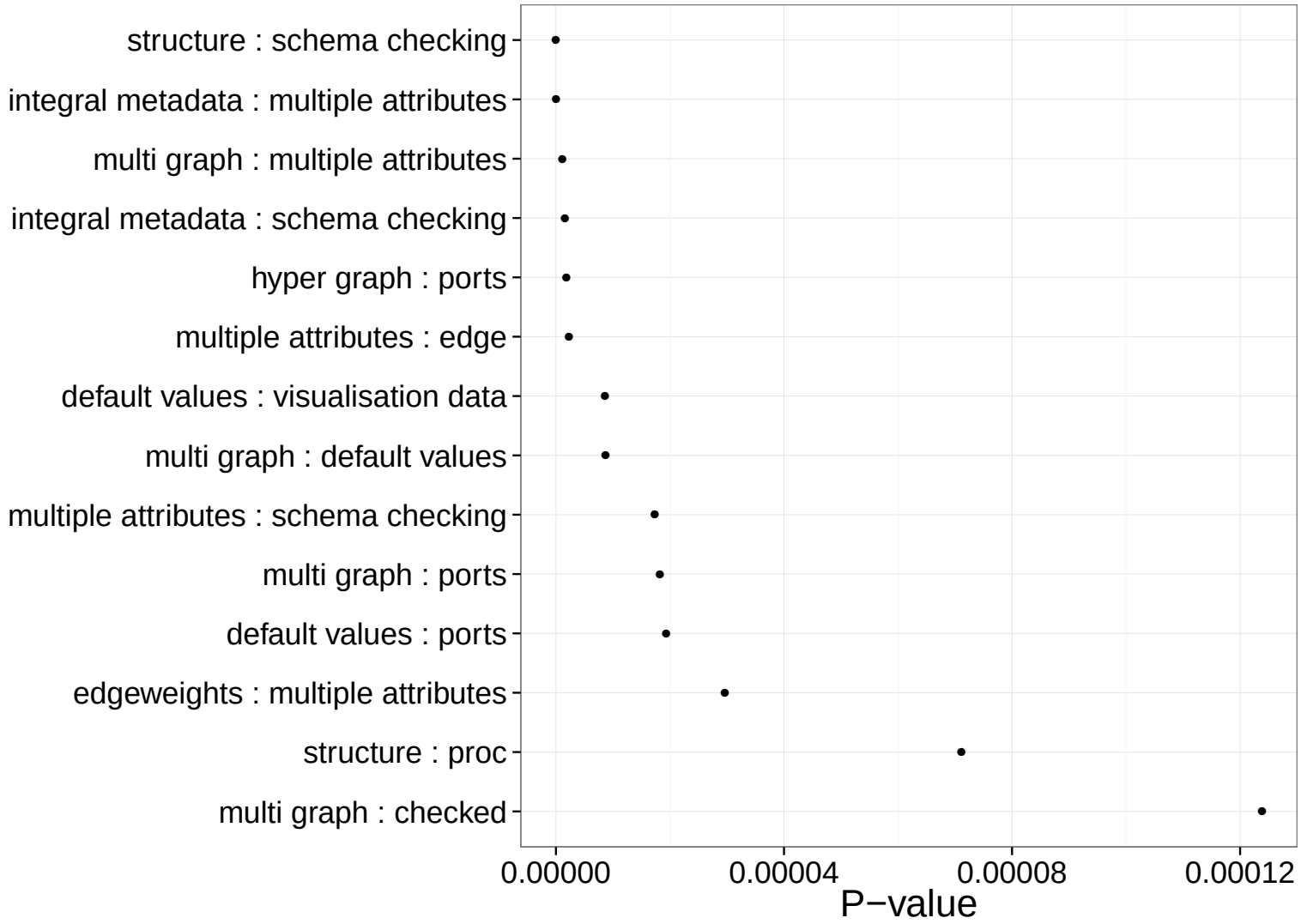
# Descriptors: General

- Extensible
- Schema checking
- Checksums
- Multiple graphs
- External references

# Common Attributes



# Attribute Correlations



# Lot's of other considerations

- Software support
  - documentation
  - maintenance
  - issue of partial support
- Public data
  - how many people already use it
- Efficiency
- Human readability
  - self-describing

# What's missing?

A lot of DB concepts

- Most are designed to be read in one piece
  - no data partitioning
  - no parallelisation
  - no serialisation
  - no random access
- Most are not designed with data curation in mind
  - no support for editing
  - no support for version, ...
- Graph DBs handle these
  - but not the portability/exchange issues

# What's missing?

Compression

- Storing large graphs
  - similar to storing images
  - nice to have native compression
- Only one (BVGraph) treats compression seriously
  - couple of others take a little care but aren't true compressions
- Most have XML-like bloat
  - file compresses OK, but read/write performance?
- There are a few papers on graph compression, but little work on formats that support it

# Conclusion

- Graph Exchange Formats
  - very many
  - lots of overlap
  - lots of variety
  - no “one true” format
- Perhaps we need three:
  - TGF (Trivial Graph Format)
  - GraphML (Feature rich, extensible, ...)
  - Format that can cope with really big graphs

Maybe a container format is what we really need?