IN33B-0125: Heterogeneity and Heterarchy: How far can network analyses in Earth and space sciences take us?

Peter Arthur Fox (pfox@cs.rpi.edu), **Ahmed Eleish** (eleisa@rpi.edu), **Congrui Li** (lic10@rpi.edu), **Mark Parsons** (parsom3@rpi.edu), **Anirudh Prabhu*** (prabha2@rpi.edu), and **Hao Zhong** (zhongh3@rpi.edu) **Rensselaer Polytechnic Institute, Tetherless World Constellation, Troy, NY, United States.**

Abstract

❖The vast majority of explorations of Earth systems are limited in their ability to effectively explore the most important (and often most difficult) problems because they are forced to interconnect at the data-element, or syntactic, level rather than at a higher scientific, or conceptual/semantic, level.

❖Recent successes in the application of complex network theory and algorithms to **mineralogy, fossils and proteins** (and the basic elements they feature) over billions of years of Earth's history, raise expectations that more general graph-based approaches offer the opportunity for new discoveries = needles instead of haystacks.

❖In the past ~ 10 years in the natural sciences there has substantial progress in providing both specialists and non-specialists the ability to describe in machine readable form, geophysical quantities and relations among them in meaningful and natural ways, effectively breaking the prior syntax barrier.

❖The corresponding open-world semantics and reasoning provide higher-level interconnections. **Data is embedded in the network!**

❖That is, semantics provided around the data structures, using open-source tools, allow for discovery at the knowledge level.

❖This presentation covers the fundamentals of data-rich network analyses for geosciences, provide illustrative examples in mineral evolution and offer future paths for consideration.

❖And then: what might be possible?

Sponsors:

DEEP CARBON OBSERVATORY

Alfred P. Sloan FOUNDATION

Poster: IN33B-0125*

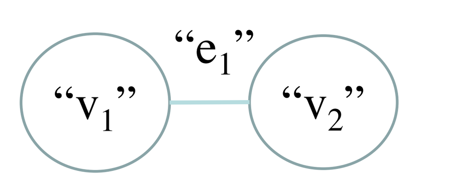
Glossary:

RPI – Rensselaer Polytechnic Institute
DCO – Deep Carbon Observatory
DTDI – Deep Time Data Infrastructure
PALEOBIODB –Paleo Biology Data Base
PROTDB – Protein Data Base

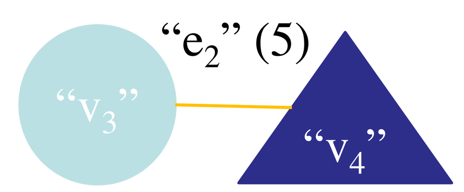
Acknowledgments:
Many thanks to the teams of researchers for the DTDI project: at Carnegie Institution for Science (Morrison, Liu et al.) for supporting domain use cases and scientific information in mineralogy; at Arizona State University for mineral data resources (RRUFF.INFO); at Rutgers (Bromberg et al.) for protein evolution; at Harvard (Muscente et al.) for fossil evolution; and to the DCO Data Science Team for their efforts of raw data extraction, algorithm extraction and general commentary: Xiaogang (Marshall) Ma, Han Wang, Benno Lee, Feifei Pan and Fang Huang.

Networks

- A network is a graph, G = (V, E)
  - V: set of labeled vertices (v_1, …v_n)
  - E: set of labeled edges (e_1, …e_n)



- Node attributes
  - Node shape, size, color
  - Can represent attributes in the data
- Link weights
- Link length, width, color, shape
- Values can be assigned to edges as weights to the relationship between 2 nodes



Metrics

Local (single node)

- How "important" is one node?
- Does one node communicate between two distinct groups?

Global (entire network)

- Is the network highly interconnected?
- Does the network form distinct groups or clusters?

Data types/ structures

NAMES	RRUFF ID	IDEAL CHEMISTRY	SOURCE	LOCALITY
Akatoreite	R060230	$Mn^{2+}_9Al_2Si_8O_{24}(OH)_8$	Michael Scott S100146	near mouth of Akatore Creek, Taieri, Otago Province, New Zealand
Akrochordite	R100028	$Mn^{2+}_5(AsO_4)_2(OH)_4 \cdot 4H_2O$	William W. Pinch	Langban, Filipstad, Varmland, Sweden
Alabandite	R070174	MnS	Michael Scott S101601	Mina Preciosa, Sangue de Cristo, Puebla, Mexico
Allactite	R070175	$Mn^{2+}_7(AsO_4)_2(OH)_8$	Michael Scott S102971	Langban, Filipstad, Varmland, Sweden
Allactite	R150120	$Mn^{2+}_7(AsO_4)_2(OH)_8$	Steven Kuitems	Sterling Mine, 1200' Level, Ogdensburg, New Jersey, USA
Alleghanyite	R060904	$Mn^{2+}_5(SiO_4)_2(OH)_2$	Michael Scott S100995	Near Bald Knob, Alleghany County, North Carolina, USA

accepted_name	accepted_attr	accepted_rank	accepted_no	early_interval	late_interval	max_ma	min_ma
Lufengpithecus	NA	genus	89270	Turolian	NA	8.7000	5.3330
Laccopithecus robustus	NA	species	58366	Turolian	NA	8.7000	5.3330
Sivapithecus indicus	NA	species	133676	Late Miocene	NA	11.6080	5.3330
Hylobatidae	NA	family	40889	Middle Pleistocene	Late Pleistocene	0.7810	0.0117
Pongo pygmaeus	NA	species	232338	Middle Pleistocene	Late Pleistocene	0.7810	0.0117
Macaca	NA	genus	40869	Middle Pleistocene	Late Pleistocene	0.7810	0.0117
Trachypithecus	NA	genus	92697	Middle Pleistocene	Late Pleistocene	0.7810	0.0117
Indraloris himalayensis	NA	species	101101	Late Miocene	NA	11.6080	5.3330
Aotus dindensis	NA	species	143115	Friasian	NA	16.3000	15.5000

source	target	value
Abenakiite-(Ce)	Adamsite-(Y)	0.83333333
Agricolaite	Albrechtschraufite	0.83333333
Adamsite-(Y)	Alexkhomyakovite	0.85714286
Adamsite-(Y)	Alstonite	0.94444444
Alexkhomyakovite	Alstonite	0.85714286
Aerinite	Alumohydrocalcite	0.96
Alloriite	Alumohydrocalcite	0.83333333
Alstonite	Alumohydrocalcite	0.97468354
Abenakiite-(Ce)	Ancylite-(Ce)	0.83333333
Adamsite-(Y)	Ancylite-(Ce)	0.83333333
Agricolaite	Ancylite-(Ce)	0.85714286
Albrechtschraufite	Ancylite-(Ce)	0.83333333
Alexkhomyakovite	Ancylite-(Ce)	0.85714286

	taxon	group
1	Aaleniella	4 (Mz)
2	Aaptochiton	6 (Pz1)
3	Aatocrinus	6 (Pz1)
4	Abactinocrinus	6 (Pz1)
5	Abadehceras	6 (Pz1)
6	Abadiella	1 (Cam)
7	Abadzekhia	2 (CzC)
8	Abakania	1 (Cam)
9	Abakanicyathus	1 (Cam)
10	Abakolia	1 (Cam)

	from	to	value
1	Aaleniella	Aaleniella	1
2	Aaleniella	Acrocoelites	1
3	Aaleniella	Acrocythere	1
4	Aaleniella	Amphitrochus	1
5	Aaleniella	Aphelocythere	1
6	Aaleniella	Bairdia	1
7	Aaleniella	Belbekella	1
8	Aaleniella	Brevibelus	1

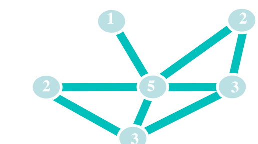
	Chalcopyrite	Malachite	Chalcocite	Bornite	Azurite	Tetrahedrite	Covellite
Chalcopyrite	25179	6949	3935	4376	3298	3654	3215
Malachite	6949	11439	2920	2437	4603	1564	2089
Chalcocite	3935	2920	5330	2468	1706	1034	1935
Bornite	4376	2437	2468	5197	1414	1094	1695
Azurite	3298	4603	1706	1414	5197	1071	1294
Tetrahedrite	3654	1564	1034	1094	1071	5010	1122
Covellite	3215	2089	1935	1695	1294	1122	3774

	X1	Abakania	Abakolia	Abdulinaspis	Abharella	Acadagnostus	Acadolenus
1	Abakania	4	0	0	0	0	0
2	Abakolia	0	3	0	0	0	0
3	Abdulinaspis	0	0	15	0	0	0
4	Abharella	0	0	0	1	0	0
5	Acadagnostus	0	0	0	0	4	1
6	Acadolenus	0	0	0	0	1	4

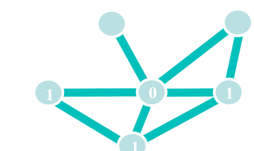
Network Analytics ~ Metrics

Metrics: Local

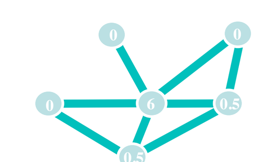
Degree is the number of links connected to a given node.



Distance is the geodesic (shortest) between any two nodes.



Betweenness is a measure of the number of geodesic paths that pass through a given node.



Metric	Earth	Mars	Moon	Vesta
Density	0.64	0.27	0.69	0.78
Maximum Network Diameter	2	3	3	2
Mean Network Diameter	1.36	1.69	1.26	1.22
Degree Central.	0.34	0.62	0.23	0.22
Betweenness Centralization	0.02	0.09	0.03	0.02

Metric	Igneous	Cr	Cu
Density	0.64	0.05	0.12
Diameter	2	6	4
Mean Distance	1.36	2.65	1.93
Degree Cent.	0.34	0.33	0.68

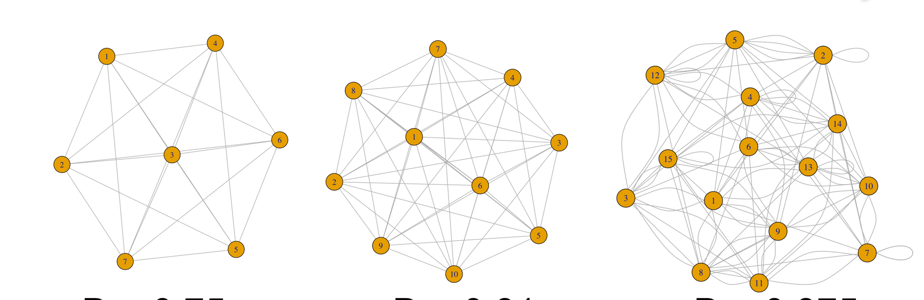
Metrics: Global

Diameter: largest geodesic distance in a network (the shortest path between the two most separated nodes)

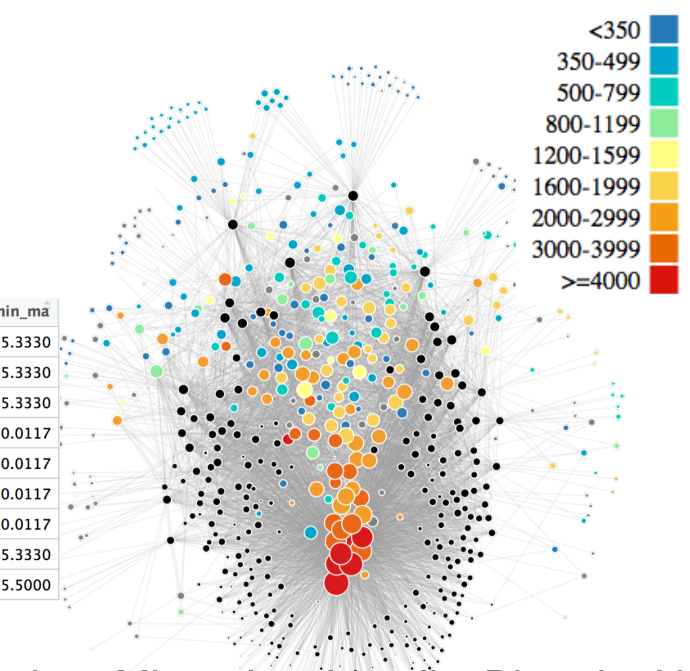
Mean Distance: average "degree of separation" in a network

Density, *D*, is the no. of links divided by the no. of possible links

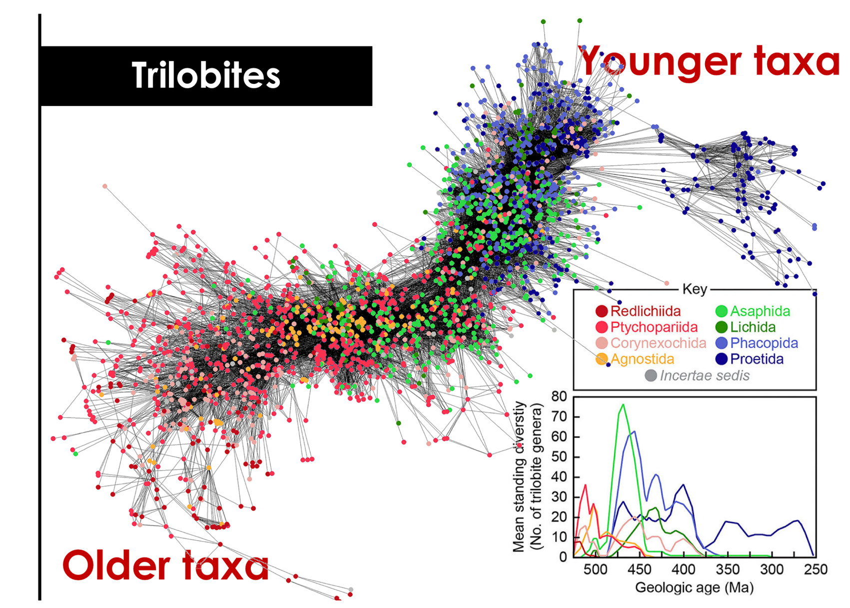
$$D = \frac{2L}{N(N-1)}$$



The networks, what they tell us



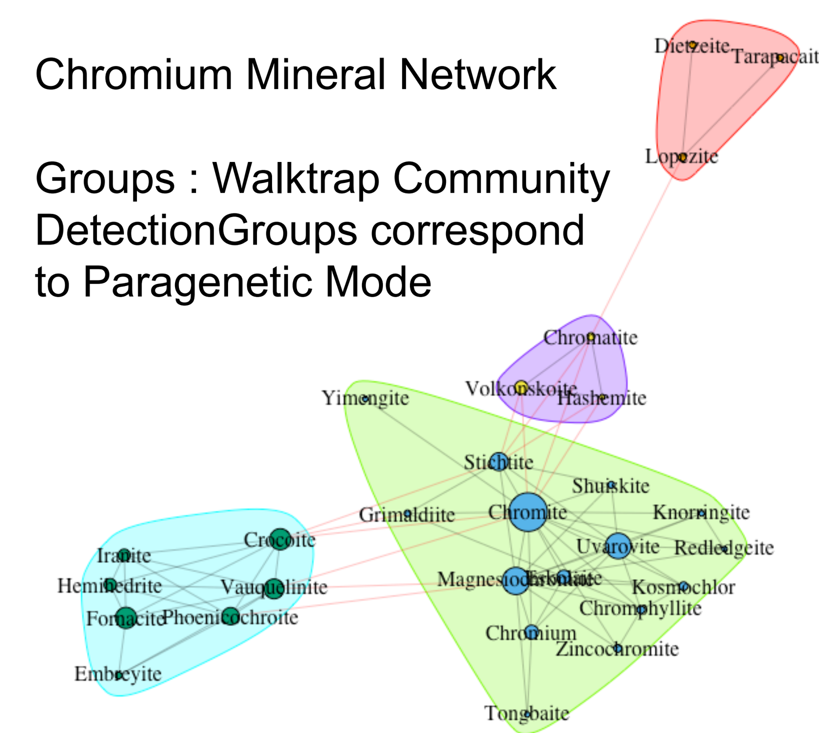
Carbon Minerals – Locality Bipartite Network

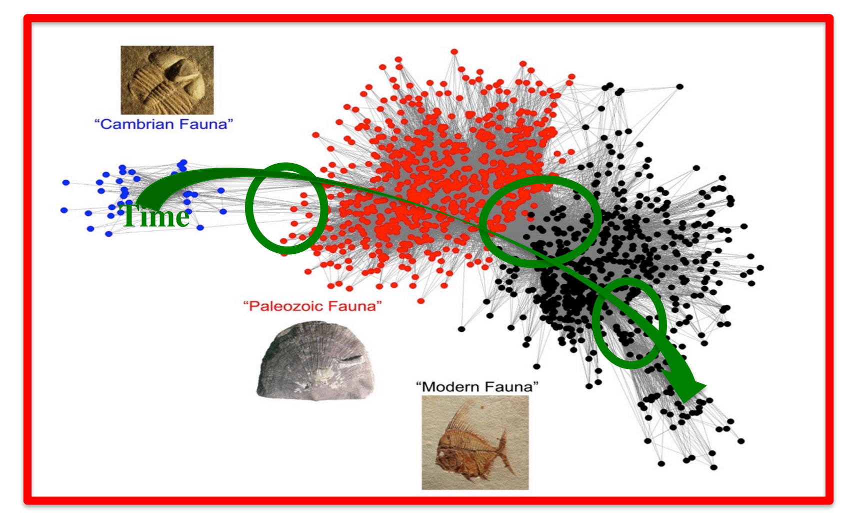


Trilobite Co-occurrence Network

Chromium Mineral Network

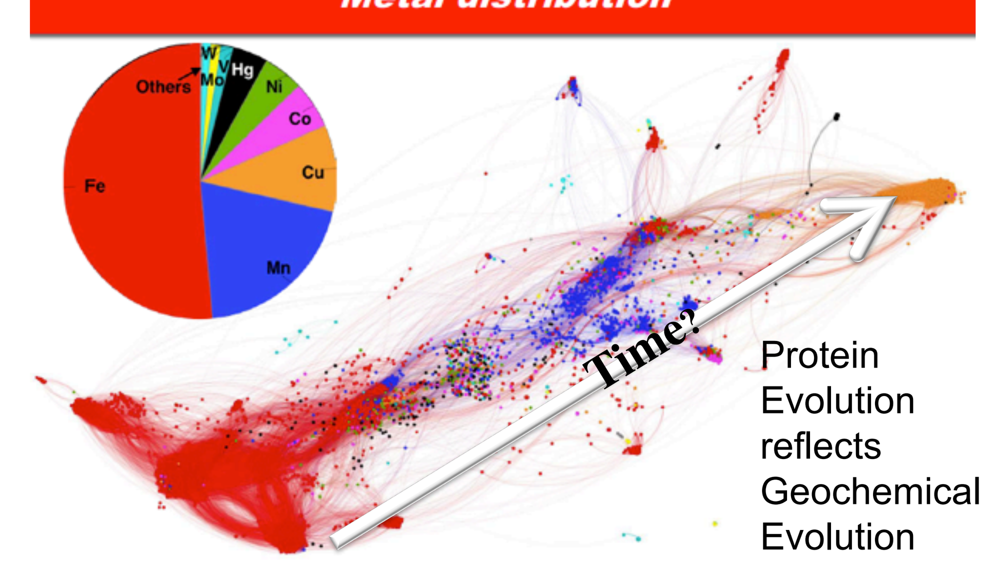
Groups : Walktrap Community DetectionGroups correspond to Paragenetic Mode





Mass Extinctions Revealed in Animal Family Networks

Metal distribution



Protein Evolution reflects Geochemical Evolution

How Far/Future work

- Network visualizations and analyses are applicable to data from many different domains.
- Network renderings of large datasets allow us to detect trends and patterns that may not be readily apparent.
- Applying network analysis methods allows us to characterize and quantify relationships across multiple dimensions.
- Future: Currently exploring: tripartite networks, 3-D networks, Virtual Reality renderings, and multi-network evolution, multiplex networks.
- How far?: Visual analyses are limited by human cognitive load; networks must be built incrementally. Need: phase space for network metrics.