NSF CHRNS CENTER FOR HIGH RESOLUTION NEUTRON SCATTERING NIST

Exploring Material Similarity using Graph-Based Crystal Structure Analysis and Machine Learning

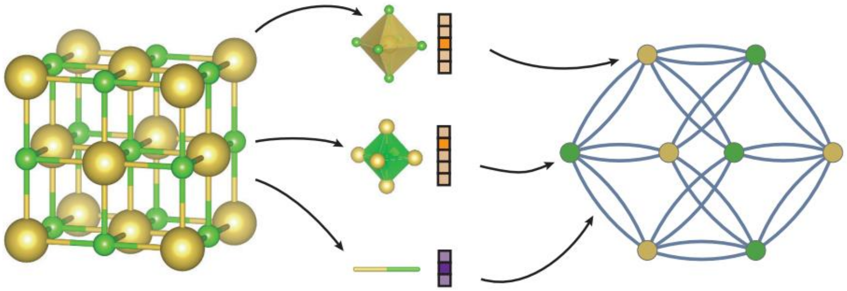
Karen Cao
Montgomery Blair High School

Mentor - Dr. William Ratcliff

NIST CENTER FOR NEUTRON RESEARCH

Objective

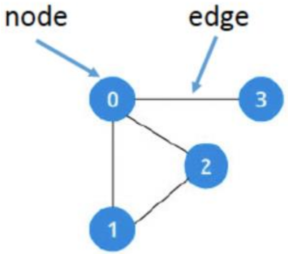
- Investigate the types of crystal structures already discovered and how they are distributed
- Represent crystal structures as graphs in order to create comparisons and place them in clusters or communities

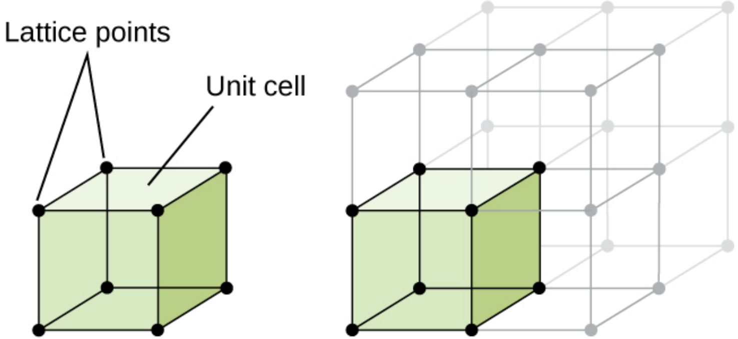


Background

What is a Crystal Structure?

- 3D arrangement of atoms, molecules, or ions in a crystalline solid
- Symmetric and repeating patterns





What is a Graph?

- Mathematical structure
- Nodes and edges

Space Groups

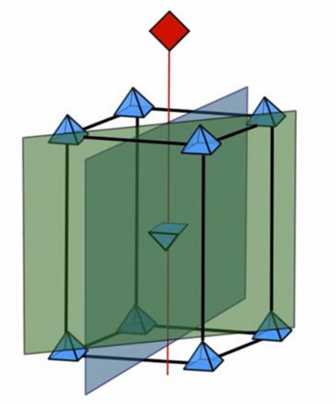
Describes arrangement of atoms in unit cell with Hermann-Mauguin (HM) symbols

Bravais types

- P – Primitive
- I – Body Centered
- F – Face Centered

Symmetry Operations

- Rotations
- Reflections
- Inversions
- Glide Plane/Screw Axis

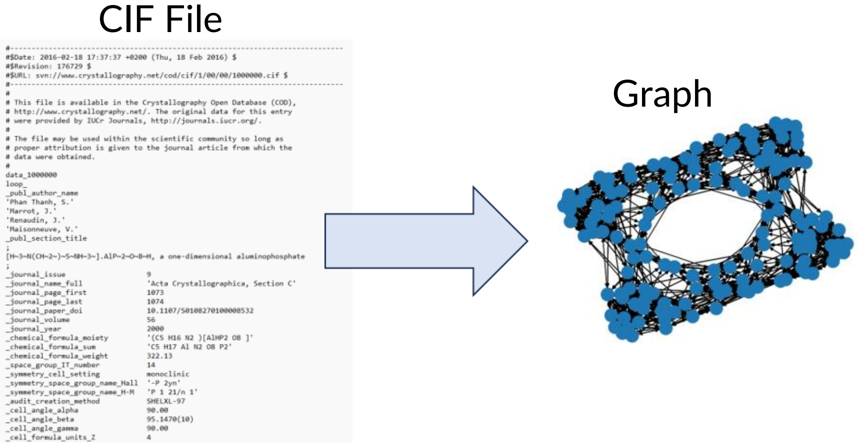


Space Group:
I 4 m m

Each space group is associated with a unique number (230 total)

Crystal Structure → Crystal Graph

- **Nearest_neighbor_edges** - k-NN (k-nearest neighbor) edge list
- **Radius_graph** - edge list based on a specified cutoff radius after several conversions and calculations

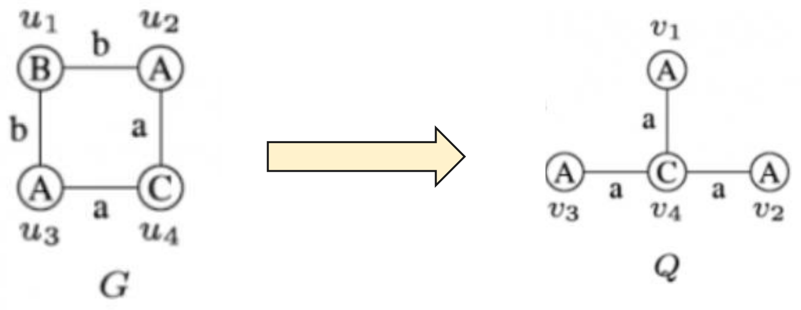


```
#------------------------------------------------------------------------------
#$Date: 2016-02-18 17:37:37 +0200 (Thu, 18 Feb 2016) $
#$Revision: 176729 $
#$URL: svn://www.crystallography.net/cod/cif/1/00/00/1000000.cif $
#------------------------------------------------------------------------------
#
# This file is available in the Crystallography Open Database (COD),
# http://www.crystallography.net/. The original data for this entry
# were provided by IUCr Journals, http://journals.iucr.org/.
#
# The file may be used within the scientific community so long as
# proper attribution is given to the journal article from which the
# data were obtained.
#
data_1000000
loop_
_publ_author_name
'Phan Thanh, S.'
'Marrot, J.'
'Renaudin, J.'
'Maisonneuve, V.'
_publ_section_title
;
[H~3~N(CH~2~)~5~NH~3~].AlP~2~O~8~H, a one-dimensional aluminophosphate
;
_journal_issue                   9
_journal_name_full               'Acta Crystallographica, Section C'
_journal_page_first              1073
_journal_page_last               1074
_journal_paper_doi               10.1107/S0108270100008532
_journal_volume                  56
_journal_year                    2000
_chemical_formula_moiety         '(C5 H16 N2 )[AlHP2 O8 ]'
_chemical_formula_sum            'C5 H17 Al N2 O8 P2'
_chemical_formula_weight         322.13
_space_group_IT_number           14
_symmetry_cell_setting           monoclinic
_symmetry_space_group_name_Hall  '-P 2yn'
_symmetry_space_group_name_H-M   'P 1 21/n 1'
_audit_creation_method           SHELXL-97
_cell_angle_alpha                90.00
_cell_angle_beta                 95.1470(10)
_cell_angle_gamma                90.00
_cell_formula_units_Z            4
```

The creators of the library to generate the crystal graph are Dr. Kamal Choudhary and Brian DeCost.

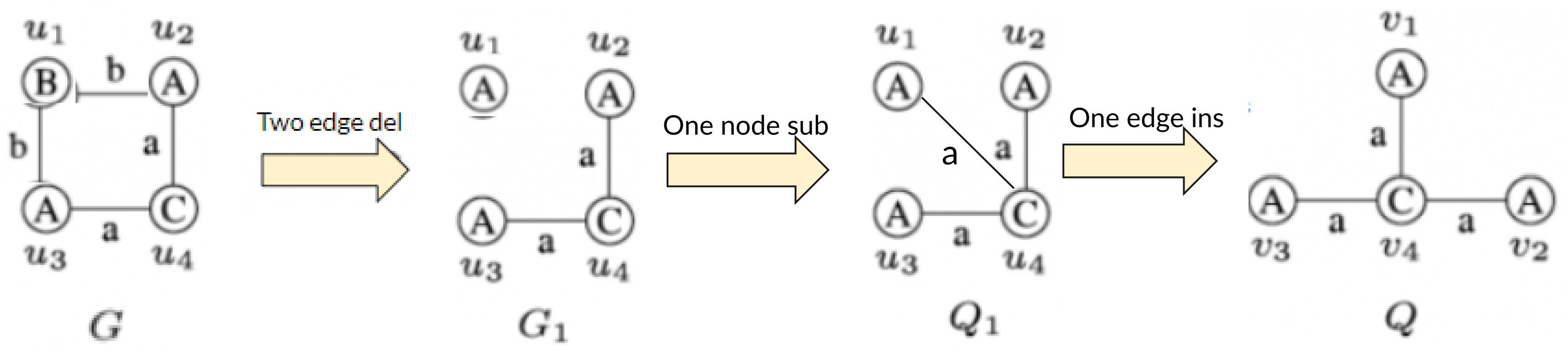
Graph Edit Distance (GED)

- Measure of similarity between two graphs
- Minimum cost to transform one graph into another through a sequence of edit operations
- Edit operations - node/edge insertion, deletion, and substitution



The idea of using graph edit distance was inspired by Dr. Debra Audus

GED Example Visualization Animation



Two edge del

One node sub

One edge ins

GED Calculation Algorithm

Based on the A* Search Algorithm[1]

1. Priority queue to store nodes and edges
2. Apply each possible operation to create a new graph state and calculate the cost function f(p)
3. Remove node with smallest f(p) from priority queue

$$f(p) = g(p) + h(p)$$

g = cumulative cost
Initial state→ current state

h = estimate cost
Current state → goal state

[1] https://hal.science/hal-01168816

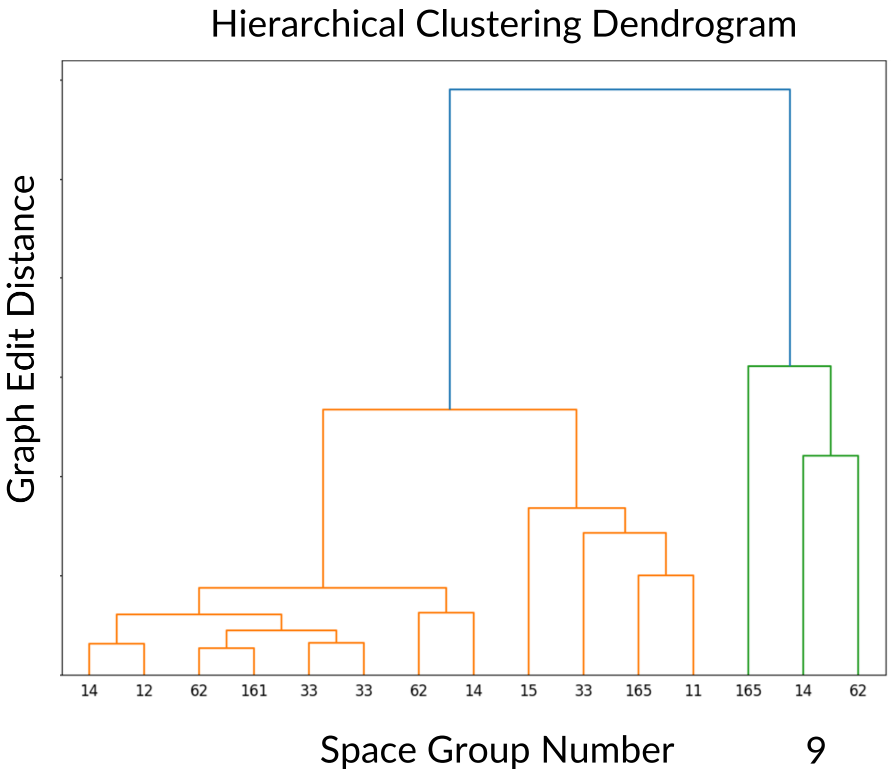
GED Calculation

graph_edit_distance(G1, G2, timeout=60)

- NetworkX Python Library
- Based on the A* Search Algorithm
- Calculates the GED between G1 and G2
- Returns the best GED calculation within a maximum number of seconds to execute

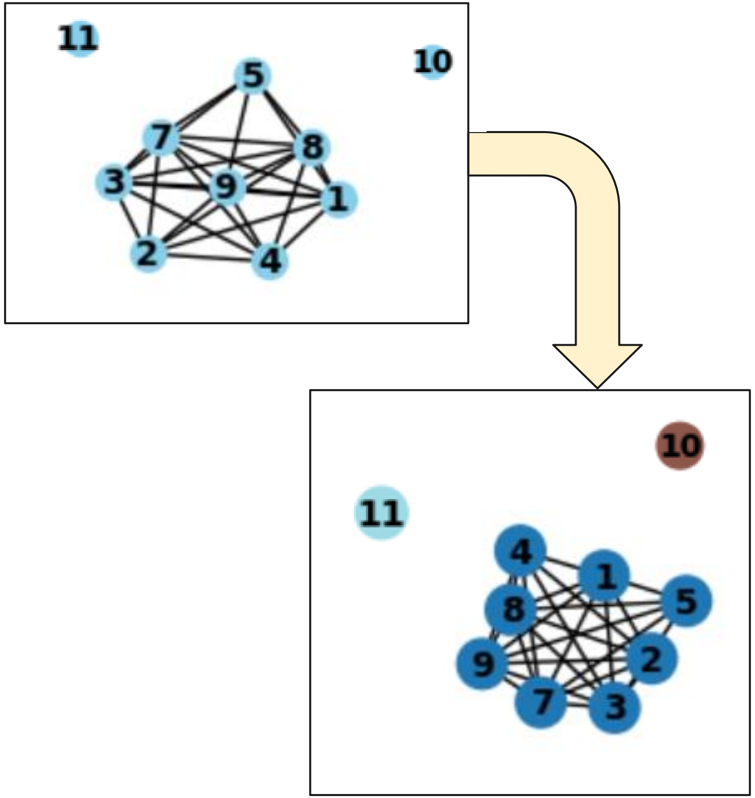
Hierarchical Clustering

- Groups similar objects into groups/clusters
- How it works:
  - Treats each graph as a separate cluster
  - Repeatedly merges two clusters that are closest
  - Iterates until all clusters merged together

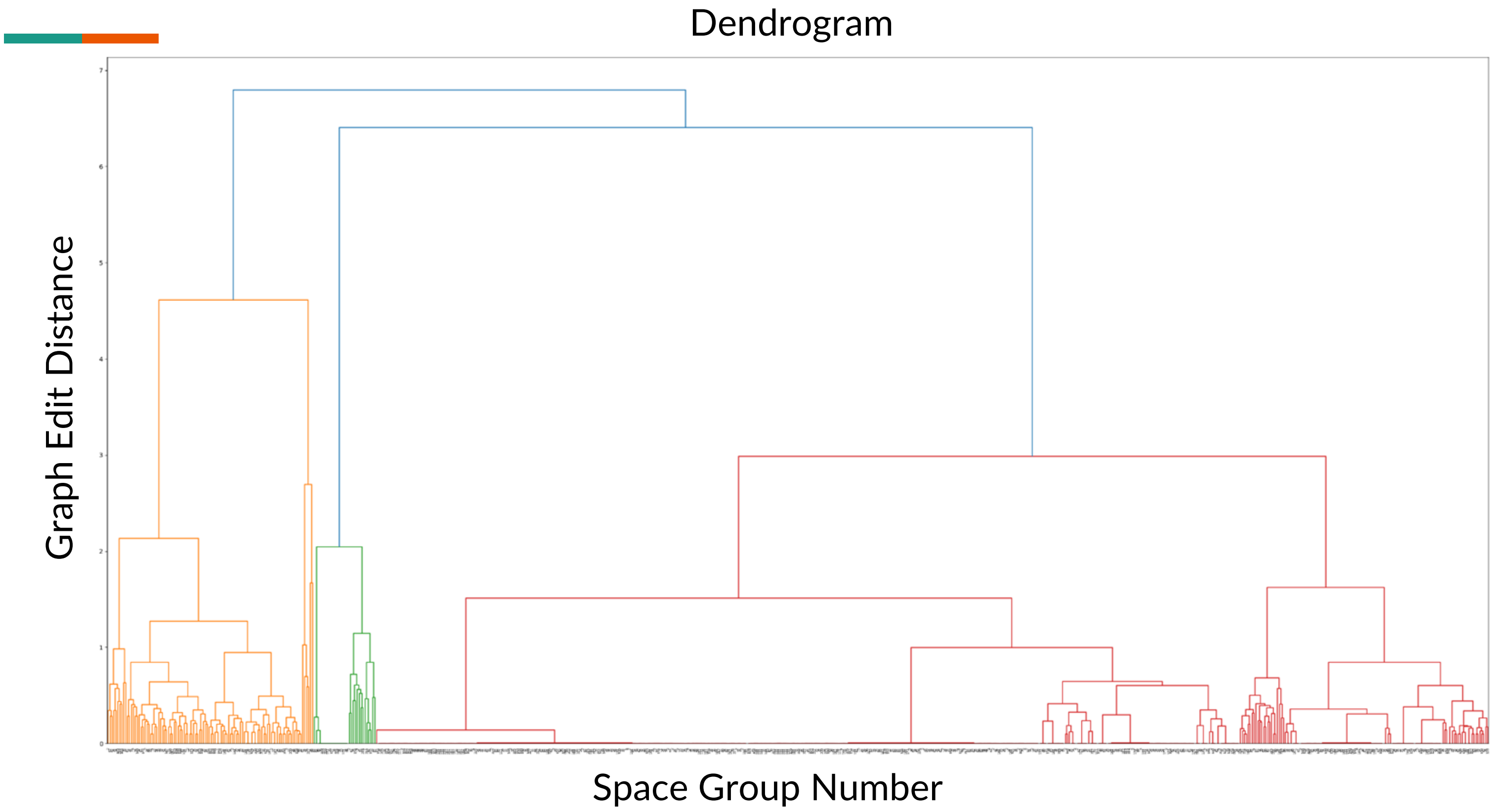


Community Detection

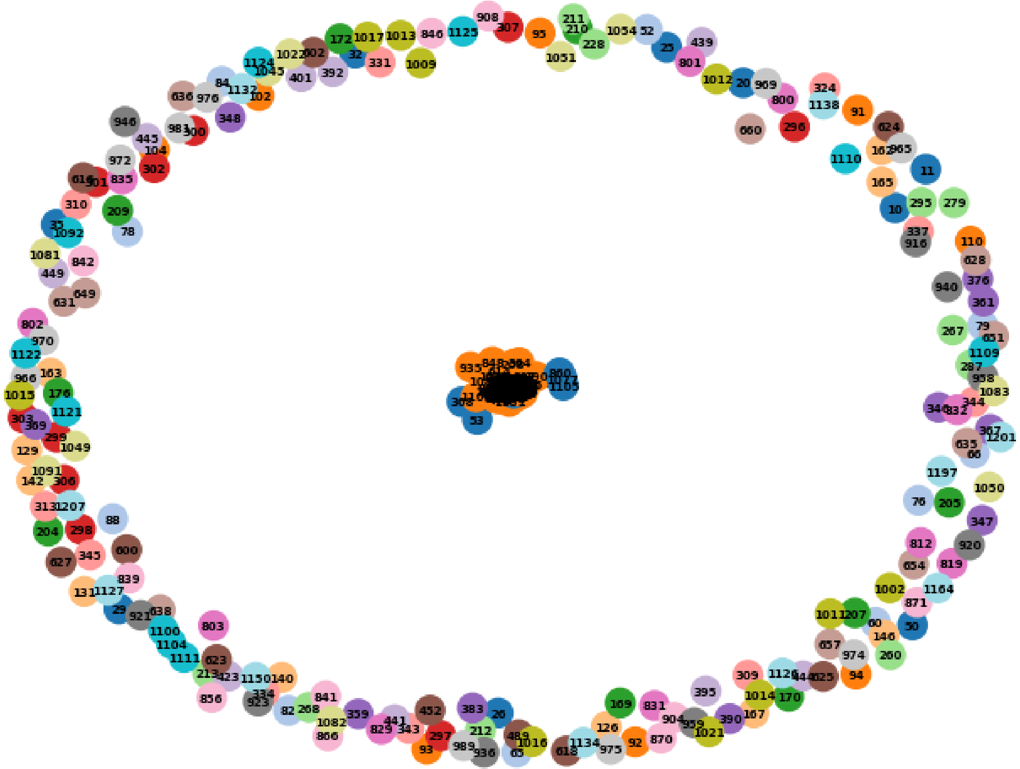
- Locates tightly connected nodes in a graph
- How it works:
  - Assigns different community to nodes
  - Considers each neighboring community for placing nodes
  - Node placed in neighbor community based on modularity

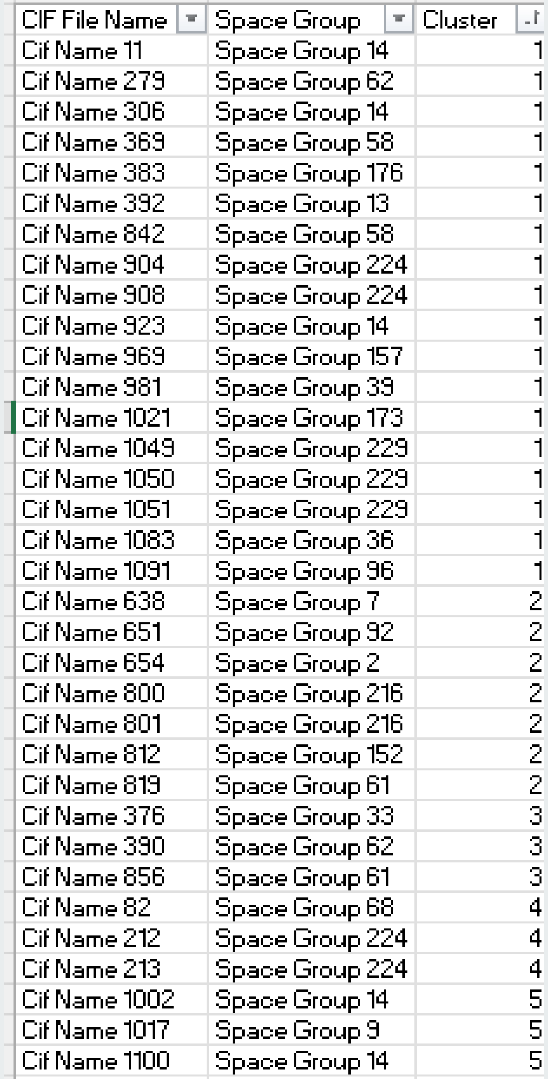


Hierarchical Clustering Results

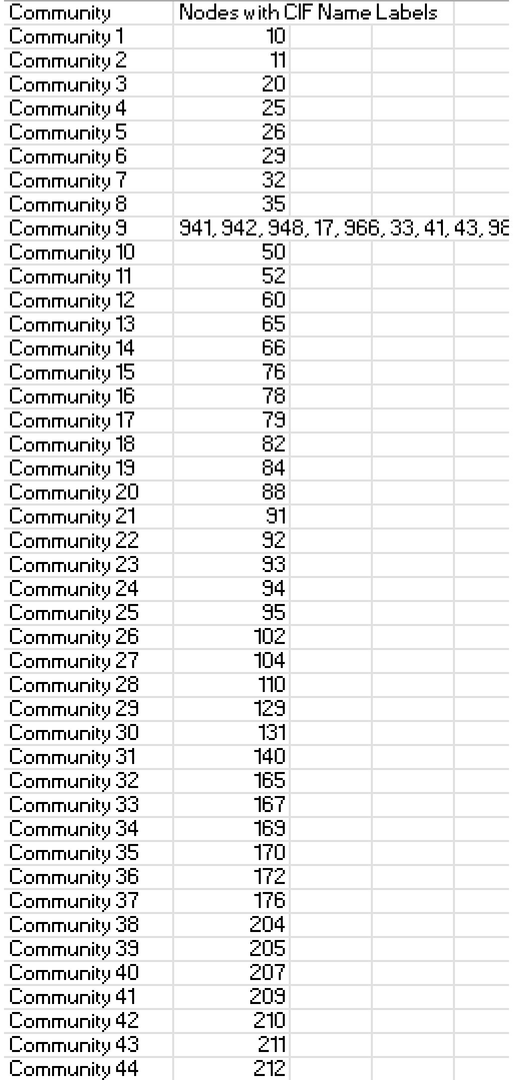


Community Detection Results





CIF File Name	Space Group	Cluster
Cif Name 11	Space Group 14	1
Cif Name 279	Space Group 62	1
Cif Name 306	Space Group 14	1
Cif Name 369	Space Group 58	1
Cif Name 383	Space Group 176	1
Cif Name 392	Space Group 13	1
Cif Name 842	Space Group 58	1
Cif Name 904	Space Group 224	1
Cif Name 908	Space Group 224	1
Cif Name 923	Space Group 14	1
Cif Name 969	Space Group 157	1
Cif Name 981	Space Group 39	1
Cif Name 1021	Space Group 173	1
Cif Name 1049	Space Group 229	1
Cif Name 1050	Space Group 229	1
Cif Name 1051	Space Group 229	1
Cif Name 1083	Space Group 36	1
Cif Name 1091	Space Group 96	1
Cif Name 638	Space Group 7	2
Cif Name 651	Space Group 92	2
Cif Name 654	Space Group 2	2
Cif Name 800	Space Group 216	2
Cif Name 801	Space Group 216	2
Cif Name 812	Space Group 152	2
Cif Name 819	Space Group 61	2
Cif Name 376	Space Group 33	3
Cif Name 390	Space Group 62	3
Cif Name 856	Space Group 61	3
Cif Name 82	Space Group 68	4
Cif Name 212	Space Group 224	4
Cif Name 213	Space Group 224	4
Cif Name 1002	Space Group 14	5
Cif Name 1017	Space Group 9	5
Cif Name 1100	Space Group 14	5



Community	Nodes with CIF Name Labels
Community 1	10
Community 2	11
Community 3	20
Community 4	25
Community 5	26
Community 6	29
Community 7	32
Community 8	35
Community 9	941, 942, 948, 17, 966, 33, 41, 43, 98
Community 10	50
Community 11	52
Community 12	60
Community 13	65
Community 14	66
Community 15	76
Community 16	78
Community 17	79
Community 18	82
Community 19	84
Community 20	88
Community 21	91
Community 22	92
Community 23	93
Community 24	94
Community 25	95
Community 26	102
Community 27	104
Community 28	110
Community 29	129
Community 30	131
Community 31	140
Community 32	165
Community 33	167
Community 34	169
Community 35	170
Community 36	172
Community 37	176
Community 38	204
Community 39	205
Community 40	207
Community 41	209
Community 42	210
Community 43	211
Community 44	212

What’s Next

- Closely investigate the clusters and communities
- How has this distribution has changed over time?
- Continue calculating the graph edit distance for a complete 200,000 x 200,000 distance matrix

Special Thanks

Dr. William Ratcliff & Mr. Paul Kienzle

Dr. Paul Butler, Mr. Jeff Krzywon, Dr. Yun Liu

Dr. Julie Borchers, Dr. Leland Harriger

Any Questions?