A Graph-Based Approach towards Discerning Inherent Structures in a Digital Library of Formal Mathematics

L. Lorigo, J. Kleinberg, R. Eaton, R. Constable

Department of Computer Science, Cornell University,
Ithaca, NY USA
{lolorigo, kleinber, eaton, rc}@cs.cornell.edu
http://www.cs.cornell.edu

Outline

- Objective & Motivation
- Methodology & Design
- Analysis of Dependency Structure in FDL & HITS Results
- Applications & Conclusion

Objective

- Organize digital formal math by **automated** and **adaptable** means
- Exploratory analysis of formal structure

Motivation

- Collections are growing large; need for better automation, search, visualization, personalization *<very broad>*
- Can we exploit it in the same was as is done with the structure of the web?
- Formal Mathematics is rich with structure; can we exploit this as some kind of metadata?

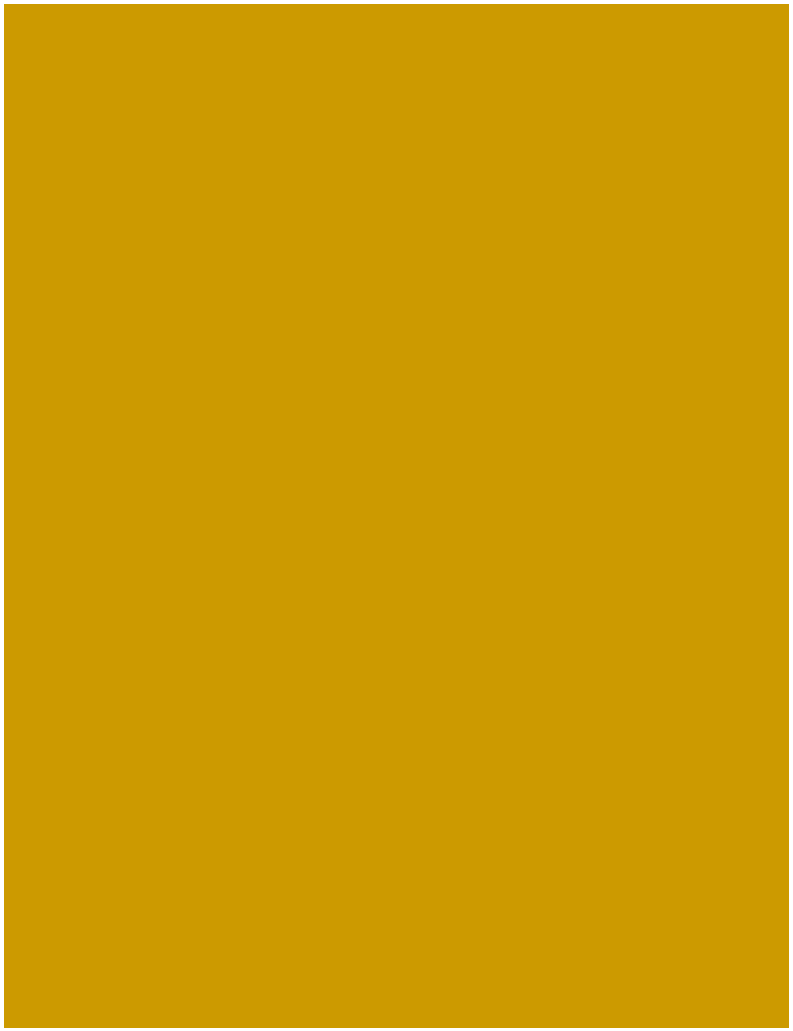
Motivation

- Collections are growing large; need for better automation, search, visualization, personalization *<very broad>*
- Can we exploit it in the same was as is done with the structure of the web?
- Formal Mathematics is rich with structure; can we exploit this as some kind of metadata?

Method

- Graph-based approach to exploit meta-information from the structure
  - relationships between theories are hidden, mass of objects, ordered by humans
- Internet search engines seems really smart; they exploit structure
  - Rank results according to “authority”, Google’s Page Rank
  - Teoma (www.teoma.com) allows you to “refine” search based on clusters
- *Equate hyperlinks with dependencies, use “HITS”*

Kleinberg's "HITS" Algorithm



- *Hubs* point to good *authorities* & *authorities* point to good *hubs*
- Sparse adjacency matrix, A
  - Hub vector = 1st eigenvector of AA^T
  - Authority vector = 1st eigenvector of A^TA
- Communities
  - Strongest hubs and authorities from respective non-principle eigenvectors
  - Currently there are improved techniques relying on the same graph

Iterative Algorithm

- Define $x^{<n>}$, $y^{<n>}$ authority, hub weight for node n in graph G with edges E.
- $x^{<n>} \leftarrow \sum_{q:(q,n)\epsilon E} y^{<q>}$
- $y^{<n>} \leftarrow \sum_{q:(n,q)\epsilon E} x^{<q>}$

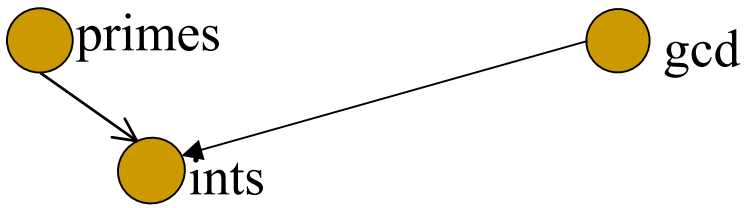
- Iterate ~ 20 times, normalizing after each.

Design Decisions

- What are analogous to hubs and authorities in formal math?
- Dependency Graphs
  - Theorems and Definitions (Rules, Tactics?)
  - Definitions have no (logical) dependencies
- Direct Forward Dependencies

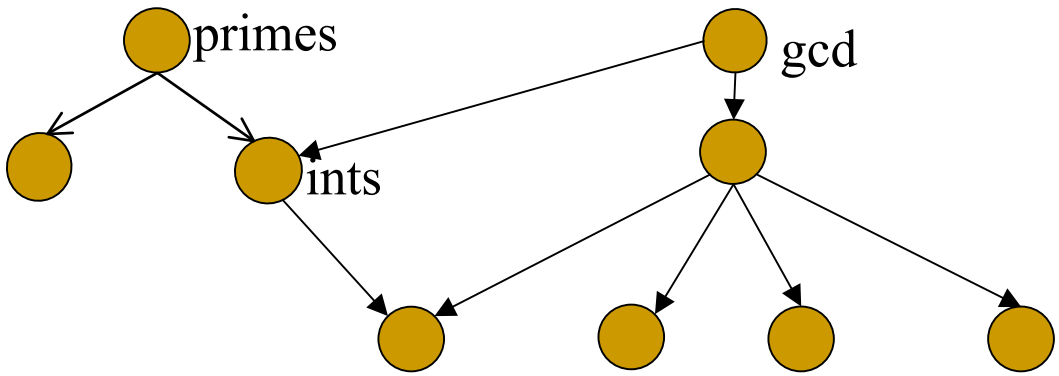
Design Decisions

- What are analogous to hubs and authorities in formal math?
- Dependency Graphs
  - Theorems and Definitions (Rules, Tactics?)
  - Definitions have no (logical) dependencies
- Direct Forward Dependencies
- Control over size and components of Graph, “Seeding”



Design Decisions

- What are analogous to hubs and authorities in formal math?
- Dependency Graphs
  - Theorems and Definitions (Rules, Tactics?)
  - Definitions have no (logical) dependencies
- Direct Forward Dependencies
- Control over size and components of Graph, “Seeding”



Implementation

- Implemented in LISP, add-on to FDL, “Formal Digital Library”
- Matlab for communities

Outline

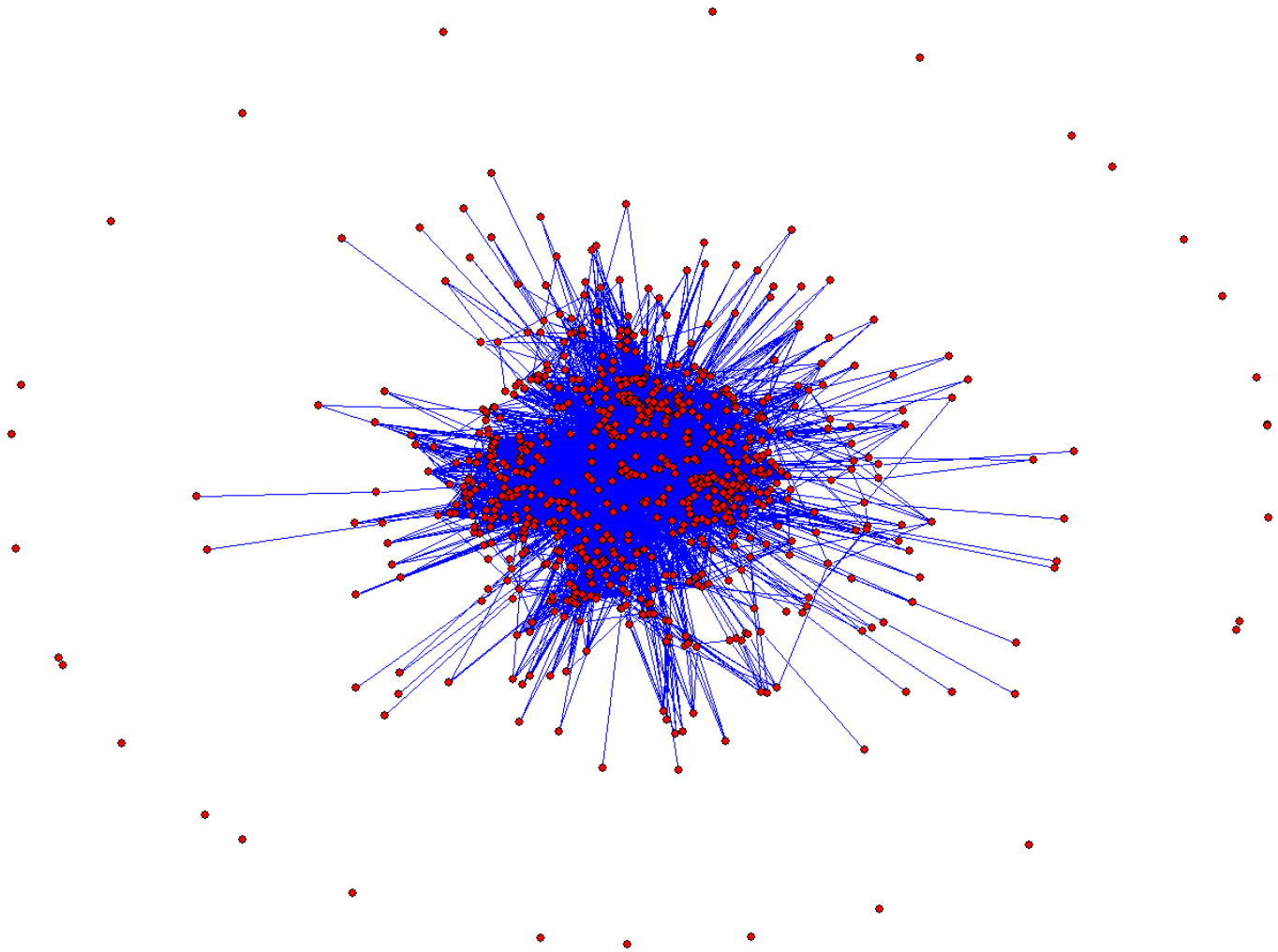
- Objective & Motivation
- Methodology & Design
- Analysis of Dependency Structure in FDL & HITS Results
- Applications & Conclusion

Graph Statistics

- Nuprl5 Standard & Num Thy Subset
- Bickford's Event Systems

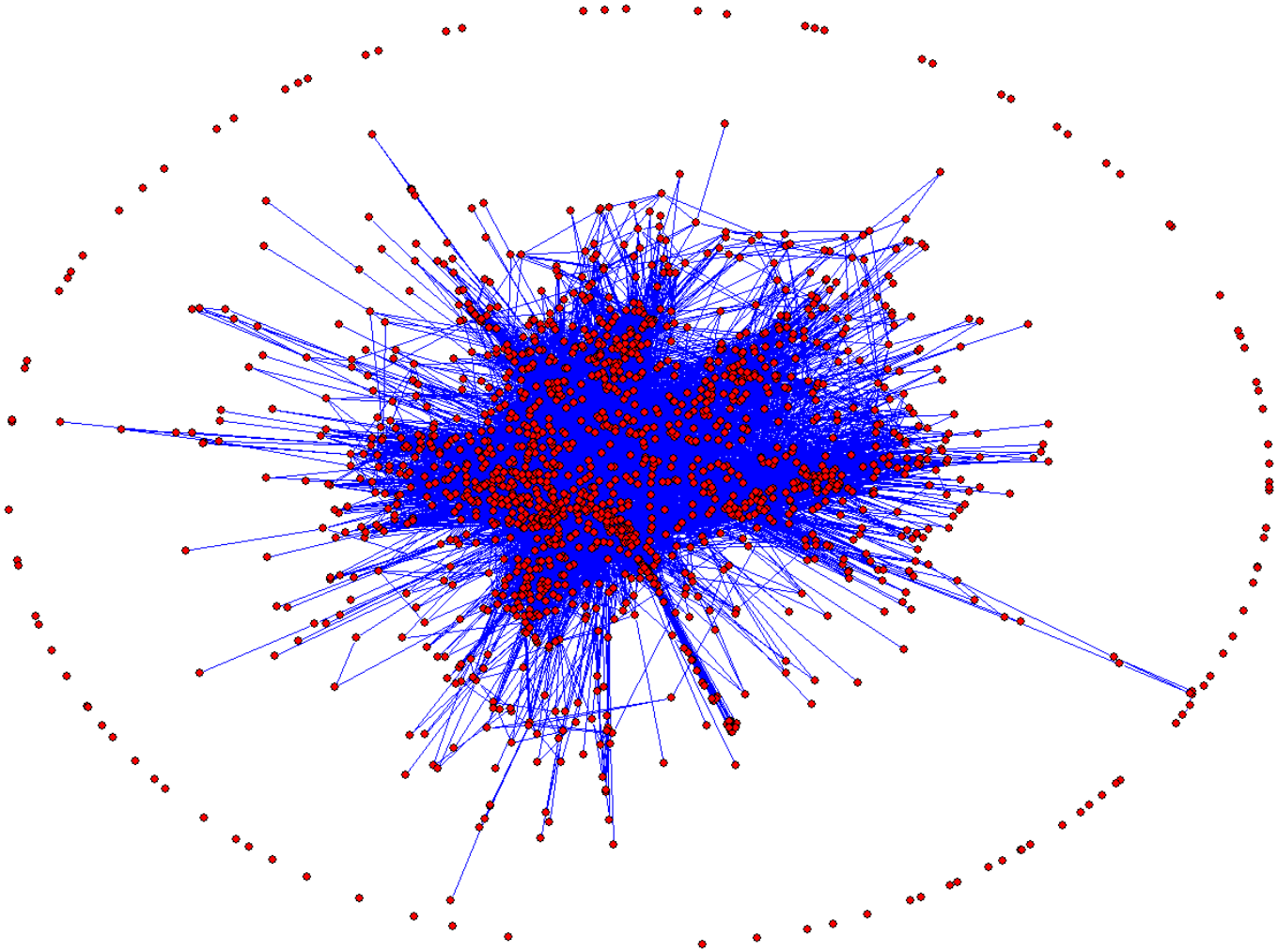
Nodes	Thm	Def	Max Outlinks	Max Inlinks	Edges	Assortativity
811	646	165	58	637	8765	-0.2949
328	260	68	56	257	3397	-0.2848
1795	1306	489	210	708	16648	-0.15896

Nuprl5 Standard



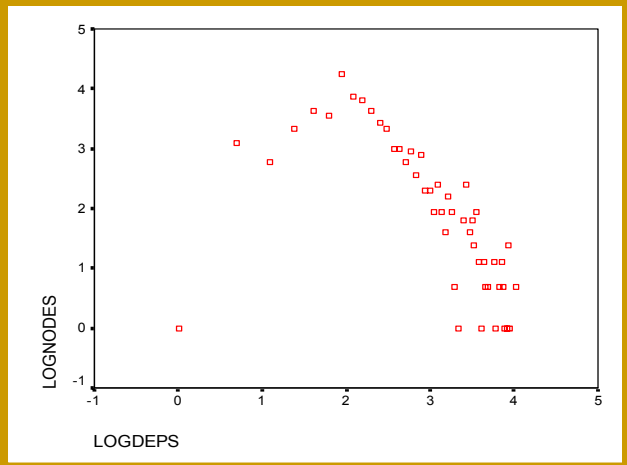
Nuprl5 Number Theory

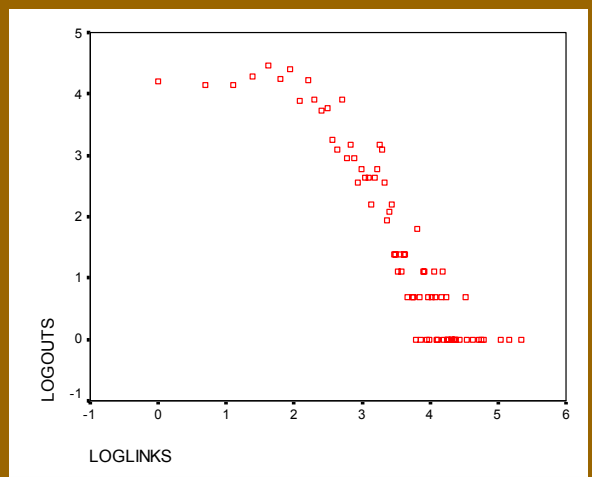
Event Systems



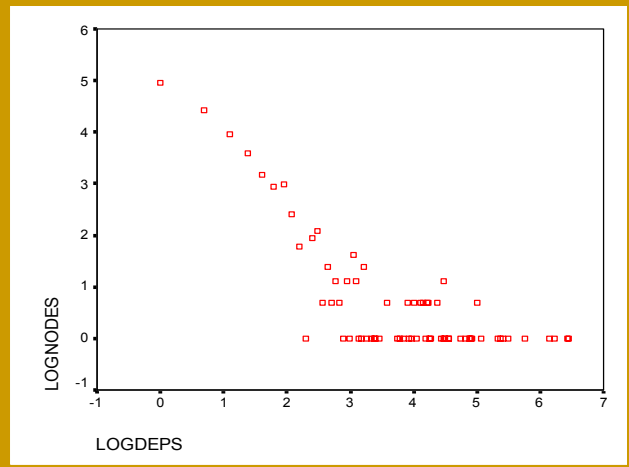
Link Distributions

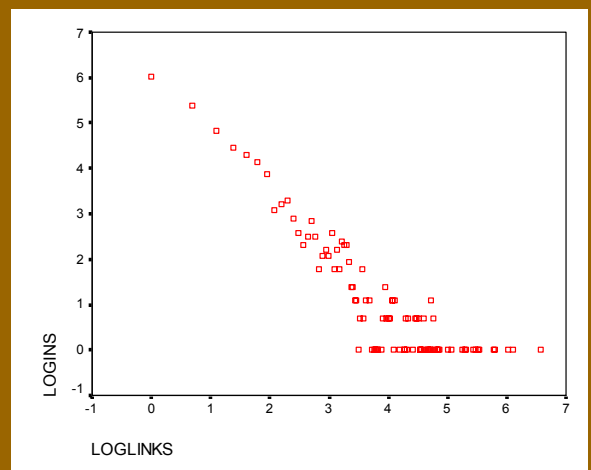
OUT LINK





IN LINK





HITS

AUTHORITIES	HUBS
member, all, prop, implies, and, iff, rev_implies, false	rem_mag_bound, select_listify_id, listify_wf, rem_eq_args_z, rem_base_case_z, select_firstn, listify_length, mod_bounds, modulus_wf
Id_wf, Knd_wf, IdLnk_wf, id-deq_wf, fpf_wf, fpf-cap_wf, Kind-deq_wf, fpf-dom_wf, fdf-trivial-subtype-top	R-compat-base, R-Feasible-Dsys, sends-rule, pre-rule, d-feasible-world, R-sends-rule, R-interface-base, R-Feasible-action, R-Dsys-base-wf

Standard and Numthy Communities

1	listify_length, select_listify_id, int_seg_ind, select_append_front, decidable__ex_int_seg, or, decidable, so_apply1
2	fincr_formation, fincr_wf, fincr_wf2, equiv_rel_functionality_wrt_iff
3	fib_coprime, gcd_sat_gcd_p, gcd_sat_pred, fib_wf, gcd_wf ycomb, not_wf
4	atomic_char, assert_of_eq_int, prime_elim, assert_of_eq_atom, le_wf

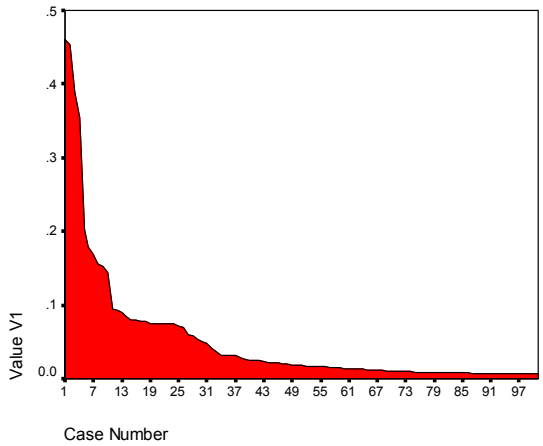
1	divides_of_absvals, absval_assoced, absval_wf
2	chrem_exists_aux_a, gcd_ex_n, chrem_exists_aux, atomic_char, prime_elim, gcd_exists_n, bezout_ident_n, chrem_exists
3	div_3_to_1, div_2_to_1, div_4_to_1, divide_wf, nequal
4	eqff_to_assert, eqtt_to_assert, assert_of_bnot, assert_of_band, prop

Event Systems Communities

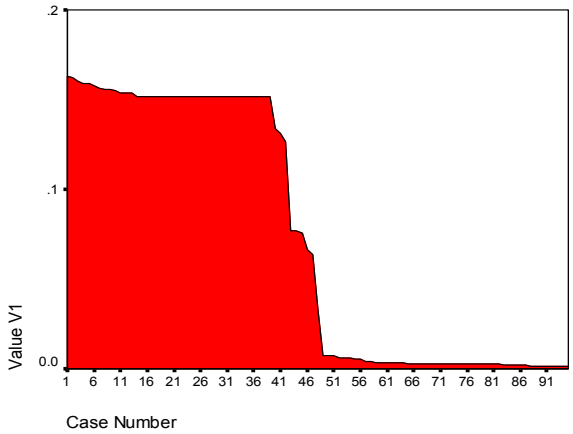
1	d-feasible-world, better-d-comp-step, d-comp-step2, d-comp-step, d-comp_wf, deq_wf
2	Rpreinit-P_wf, Rpreinit-init_wf, Rpreinit-ds_wf, Rpreinit-T_wf, Rpreinit-loc_wf, Rpreinit-a_wf, Rpreinit?, Rpreinit?_wf, Rpreinit-ds, Rpreinit
3	Reffect-f_wf, Reffect-ds_wf, Reffect-x_wf, Reffect-T_wf, Reffect-loc_wf, Reffect-knd_wf, Reffect?, Reffect?_wf
4	Rframe-loc_wf, Rframe-T_wf, Rframe-L_wf, Rframe-x_wf, Rframe?, Rframe?_wf
5	l_contains_disjoint, l_contains_append3, l_contains_append2, l_contains_append, l_contains_wf, l_contains-append4, l_contains

Weight Distributions Nuprl5

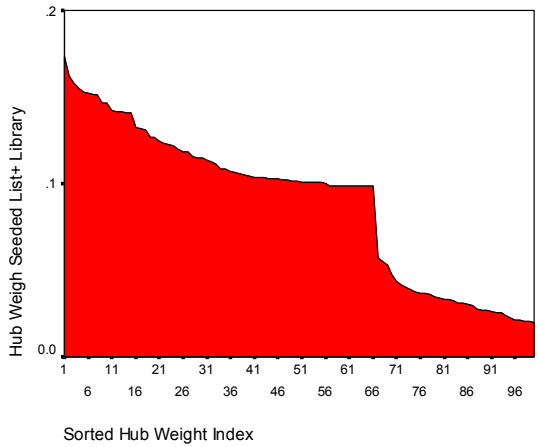
AUTHORITIES



HUBS



HUBS*

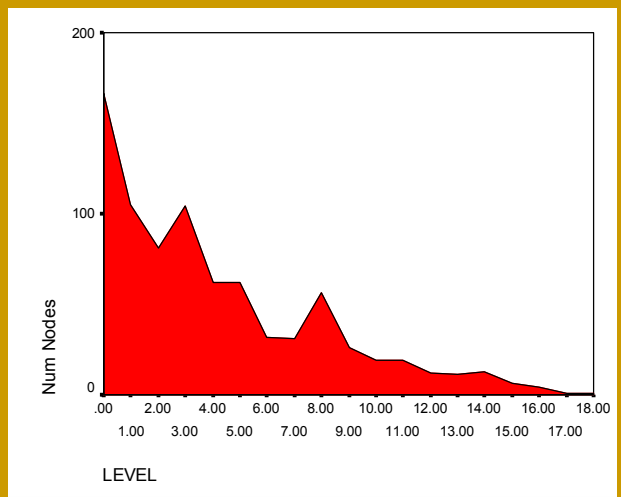


- Step-like Weight Distributions
  - Booleans, lower plateau on authority graph
  - Stratified development of hubs, Bickford's extended list theory

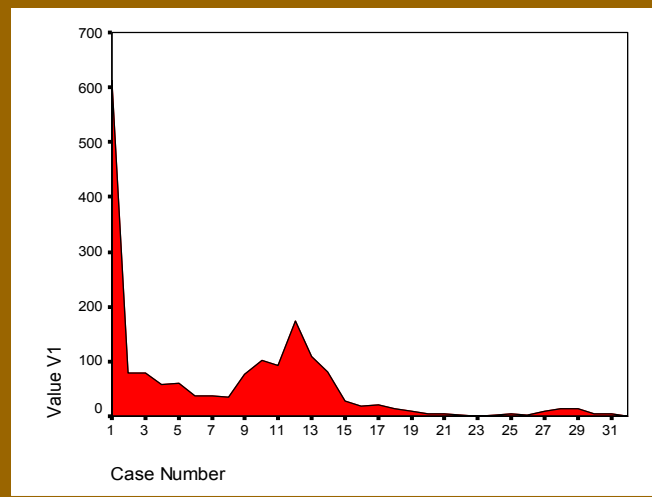
Level Distribution

- Def: A node, *n*, is in level, *i*, iff *n* depends only on nodes in levels < *i*, and *i* is the smallest value for which this is the case.
- Characteristic peaks found, suggest characteristic size/complexity of proofs

NUPRL5



EVENT SYSTEMS



Contintued Experiments

- Are the properties specific to the FDL?

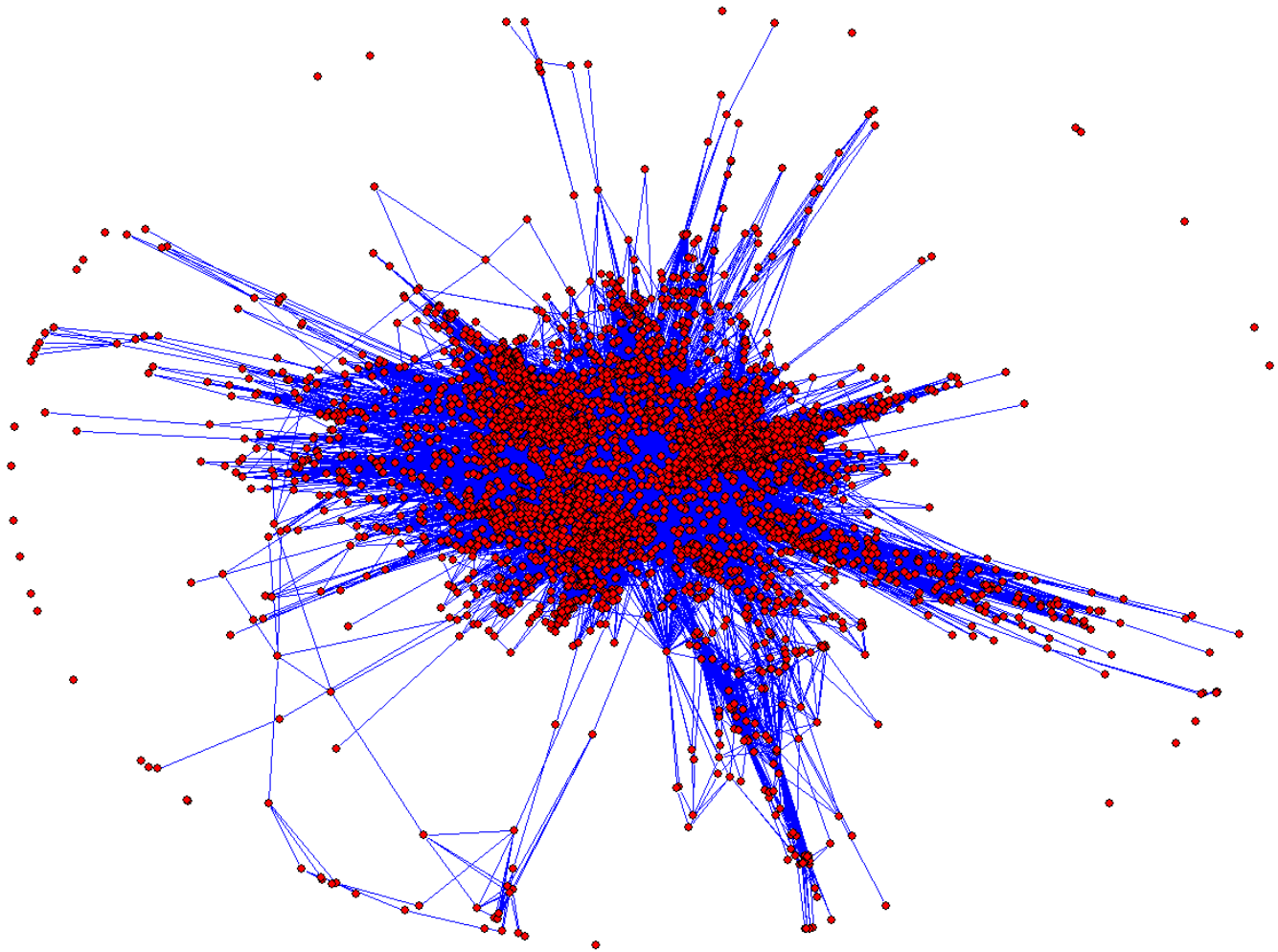
HELM Coq Library Comparison

- Nuprl5 Event Systems
- Coq in HELM

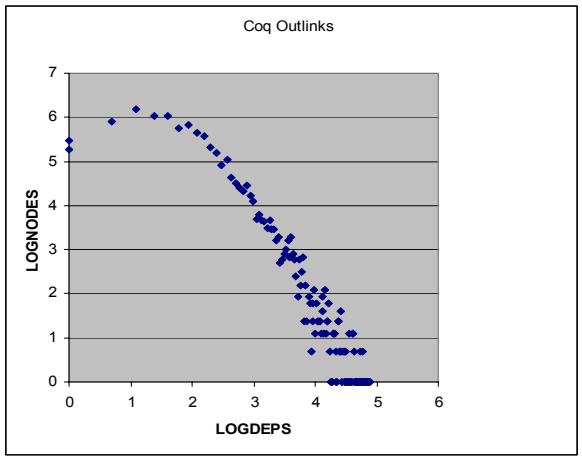
Nodes	Thm	Def	Max Outlinks	Max Inlinks	Edges	Assortativity
811	646	165	58	637	8765	-0.2949
1795	1306	489	210	708	16648	-0.15896
5346	5170	176	132	3093	63093	-0.1884

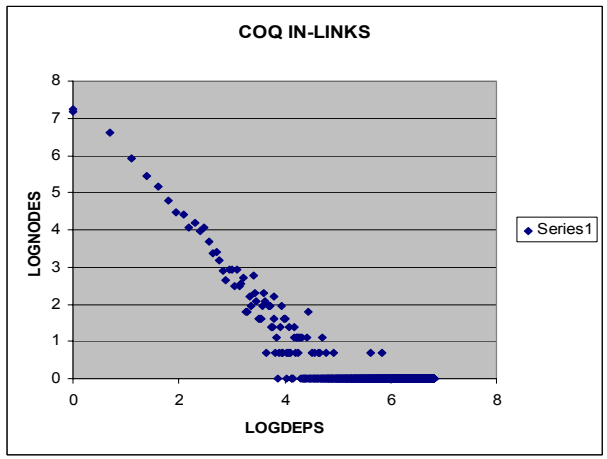
132 was /Coq/Reals/RiemannInt/RiemannInt_P28

Coq Graph

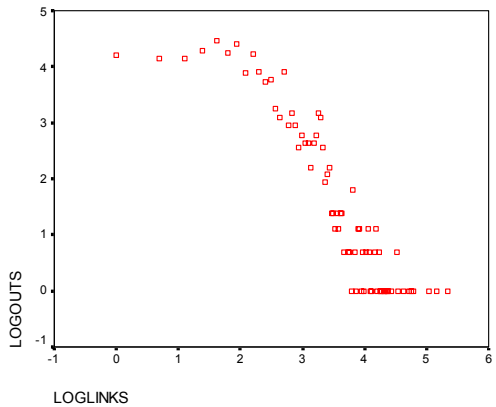


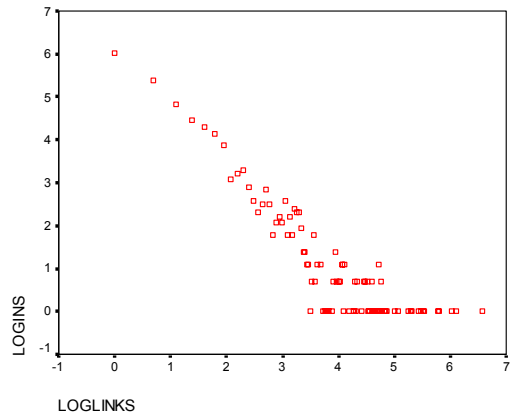
Coq Link Analysis





- Similar link distribution pattern





Coq HITS

AUTHORITIES	"HUBS"
/Init/Logic/eq.ind /Init/Logic/eq_ind /Reals/Rdefinitions/R /Init/Logic/eq_ind_r /Reals/Rdefinitions/R0 /Init/Datatypes/bool.ind /Reals/Rdefinitions/Rplus /Reals/Rdefinitions/R1 /Reals/Rdefinitions/Rmult	/Reals/Rtrigo_alt/cos_bound /Reals/SeqProp/ cv_speed_pow_fact /Reals/RiemannInt/ RiemannInt_P12 RiemannInt_P6 RiemannInt_P28 /Reals/Rtopology/Heine /Reals/Alembert/Alembert_C2 /Reals/RiemannInt/ RiemannInt_P25 RiemannInt_P8

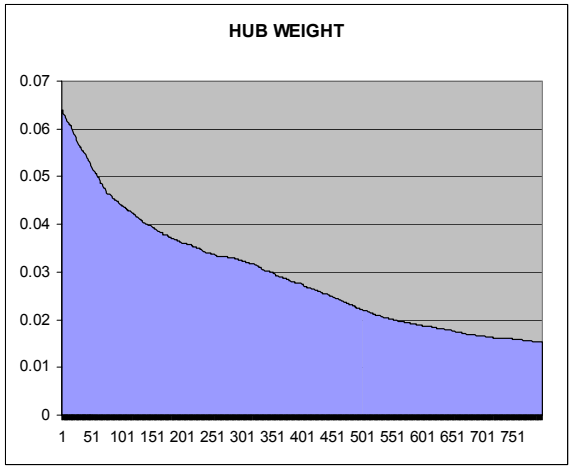
Coq Communities

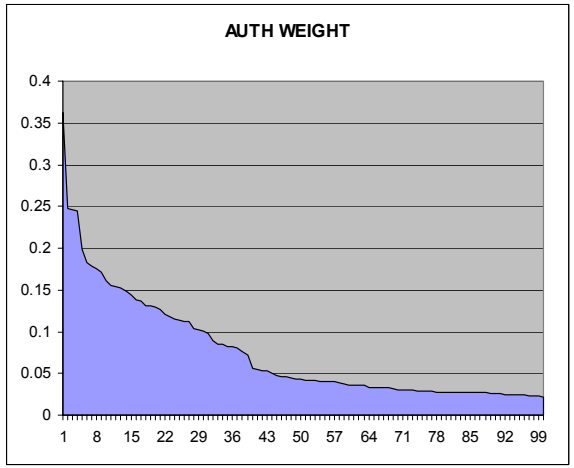
- Logic/Berardi/inv Arith/Wf_nat/induction_gtof1 Arith/Wf_nat/lt_wf_rec1 IntMap/Mapaxioms/MapMerge_assoc Logic/Berardi/i Logic/Berardi/j Bool/Bvector/Vconst
- Bool/Bvector/vector_rec Bool/Bvector/vector_ind Init/Datatypes/bool_ind Init/Datatypes/bool_rec
- Logic/ClassicalFacts/f1_o_f2 IntMap/Mapaxioms/FSetUnion_idempotent IntMap/Mapc/MapMerge_idempotent_c Logic/Diaconescu/proof_irrel Logic/Berardi/retract_rect IntMap/Mapaxioms/MapMerge_assoc Logic/Eqdep/eq_dep_dep1 Logic/Eqdep/eq_dep1_eq
- Relations/Newman/Newman Init/Datatypes/bool_rect Bool/Bvector/vector_rect IntMap/Mapaxioms/MapMerge_assoc

Coq Communities II

- Logic/Hurkens/lemma2 Logic/Hurkens/U Logic/Hurkens/sb Logic/Hurkens/le Logic/Hurkens/U
- Init/Logic_Type/identity_rect_r Init/Logic_Type/identity_rec_r Init/Logic_Type/identity_ind_r Init/Datatypes/identity_rect
- romega/ReflOmegaCore/Tred_factor3_stable omega/OmegaLemmas/fast_Zred_factor3 ZArith/auxiliary/Zred_factor3 romega/ReflOmegaCore/Tred_factor3
- ring/Ring_abstract/minus_varlist_insert_ok ring/Ring_abstract/minus_varlist_insert ring/Ring_theory/Th_plus_zero_left2 ring/Ring_theory/Th_plus_comm

Coq HITS





- Smoother weight degradation
- Very small hub values

Outline

- Objective & Motivation
- Methodology & Design
- Analysis of Dependency Structure in FDL & HITS Results
- Applications & Conclusion

Applications

- Authority weights may be used to rank search results
- Communities will be used to aid users in posting FDL material for presentation
- Overall structure is useful in comparing different collections
- Network visualizations found useful to some users

Conclusion

- Automated means can be used to discern hub, authority, and some community structure, some variations between Coq data and Nuprl
- Findings of characteristic breadth (link-degree) for bothe systems and depth (level) of the library, the latter for FDL
- Dependencies hold useful information for mkm

Acknowledgments

M. Bickford & S. Allen http://www.nuprl.org
HELM http://helm.cs.unibo.it & Coq
Pajek for layout of graph
NSF Grant #0333526
ONR Grant #N00014–01–1–0765

Thank you MKM.

lolorigo@cs.cornell.edu