UNIVERSITY OF CAMBRIDGE

A Formalisation of the Balog–Szemerédi–Gowers Theorem in Isabelle/HOL

Angeliki Koutsoukou-Argyraki | Mantas Bakšys | Chelsea Edmonds

ak2110@cam.ac.uk | mb2412@cam.ac.uk | cle47@cl.cam.ac.uk

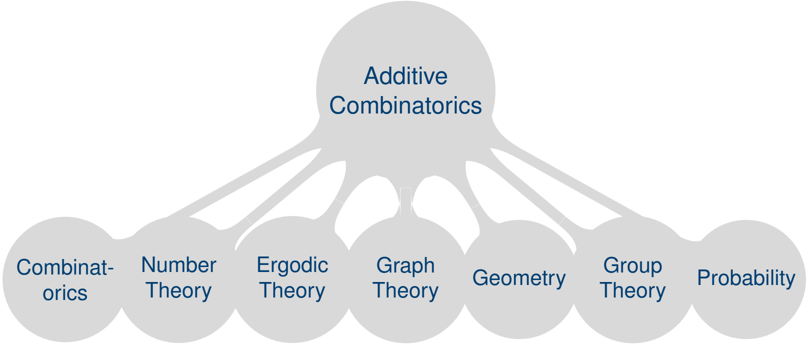
Department of Computer Science & Technology, University of Cambridge, UK

CPP2023

11 August 2023

Additive Combinatorics

Additive combinatorics is, at heart, the study of combinatorial questions involving *the additive structure of sets*



What is "Additive Structure"?

Given an additive abelian group G and finite subsets A and B we define:

- **Sumset**: $A + B = \{a + b | a \in A, b \in B\}$.
- **Difference Set**: $A - B = \{a - b | a \in A, b \in B\}$.
- **Additive Energy**:
  $E(A) = |\{(a, b, c, d) \in A \times A \times A \times A \; | a + b = c + d\}|/|A|^3$.

Simple concepts = many questions

e.g. Is the sumset a subgroup or is it close to being one? What are the bounds on cardinality? Do sets contain an arithmetic progression?

…

The Balog–Szemerédi–Gowers Theorem

Balog & Szemerédi (1994): Every finite subset A (of given additive energy) in an abelian group *must* contain a subset A' of A so that the cardinality of A' is large *but* the cardinality of the sumset $A' + A'$ is small.

Gowers (2001): New proof of the above with much better bounds on the cardinalities (BSG). Key ingredient of Gowers's new proof of the celebrated Szemerédi's Theorem on arithmetic progressions.

Why Formalise?

Modern   Interesting   Relevant

- Additive combinatorics: active field with many open problems.
- BSG formal theorem: first formalisation & crucial tool in current research on additive combinatorics.
- BSG proof techniques: a fascinating interplay between *additive combinatorics*, *graph theory* and *probability theory*.
- Formal combinatorics libraries are quickly growing. Isabelle AFP: 58/316 Mathematics entries. Mirrored in other proof assistants (e.g. Lean).

Our Key Contributions

- Background formalisations in additive combinatorics and probability theory as required.
- A new, extensive, extensible, generalised graph theory library.
- **Formalisation of the Balog–Szemerédi–Gowers Theorem in Isabelle/HOL (first in any proof assistant)**.
- A case study on the locale-centric approach in Isabelle/HOL - including the interplay in proof of different mathematical fields.
- Formalisation of some additional related results in additive combinatorics & an alternative version of the main theorem (sumsets vs difference sets).

Background Resources

- Previous work on basic sumset theory from the AFP entry "The Plünnecke-Ruzsa Inequality" by Koutsoukou-Argyraki & Paulson.
- ...built on the AFP entry "A Case Study in Basic Algebra" by Ballarin.
- 2022 lecture notes by Gowers: "Introduction to Additive Combinatorics" for the University of Cambridge.

A new general undirected graph theory library

Motivated by:

- Limitations and inflexibility of existing libraries specific to entries.
- Unnecessary complexity introduced by general *directed* library in Isabelle (Noschinski, 2015).

```
type_synonym uvert = nat
type_synonym uedge = "nat set"
type_synonym ugraph = "uvert set × uedge set"
```

Basic Undirected Graphs (Noschinski)

```
record ('v,'w) graph =
  nodes :: "'v set"
  edges :: "('v × 'w × 'v) set"
```

Dijkstra's Algorithm (Nordhoff & Lammich)

```
record ('a,'b) pre_digraph =
  verts :: "'a set"
  arcs :: "'b set"
  tail :: "'b ⇒ 'a"
  head :: "'b ⇒ 'a"
```

Digraphs (Noschinski)

Locales

Parameters

Assumptions

```
locale graph_system =
  fixes vertices :: "'a set" ("V")
  fixes edges :: "'a edge set" ("E")
  assumes wellformed: "e ∈ E ⟹ e ⊆ V"
```

```
locale ulgraph = graph_system +
  assumes edge_size: "e ∈ E ⟹
       card e > 0 ∧ card e ≤ 2"
```

```
locale sgraph = graph_system +
  assumes two_edges: "e ∈ E ⟹ card e = 2"
```

```
locale bipartite_graph = graph_system +
  fixes X Y :: "'a set"
  assumes partition: "partition_on V {X, Y}"
  assumes ne: "X ≠ Y"
  assumes edge_betw: "e ∈ E ⟹ e ∈ all_bi_edges X Y"
```

Indirect Inheritance via sublocales

```
sublocale bipartite_graph ⊆ sgraph
  using card_edges_two by (unfold_locales)
```

The Balog–Szemerédi–Gowers Theorem

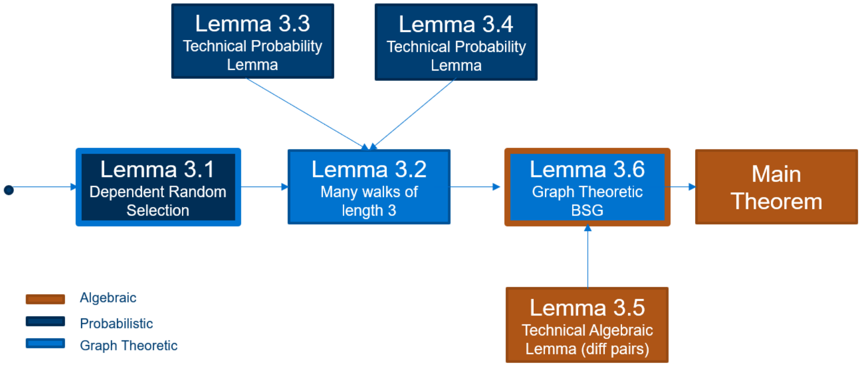
The formal statement of the theorem:

Theorem (Balog–Szemerédi–Gowers)

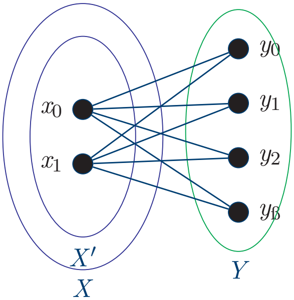
Let $c > 0$. For every finite subset of an abelian group A with additive energy $2c$ there exists a subset A' in A of cardinality at least $c^2|A|/4$ such that $|A' - A'| \leq 2^{30}|A|/c^{34}$.

Note: analogous version for sumsets, various different versions and refinements of the theorem/proof are available (Zhao, Sudakov & Szemerédi & Vu).

Sketch of the proof



The Dependent Random Selection Method



Lemma (3.1)

Intuitively: Given a dense bipartite graph ($\delta = |E|/|X||Y|$), we can restrict one of its vertex sets to a large subset in which almost all pairs of vertices are joined by many paths of length two (codegree = number of paths of length two).

How do we find a random subset with better properties, based on the structure of the original set?

The Dependent Random Selection Method

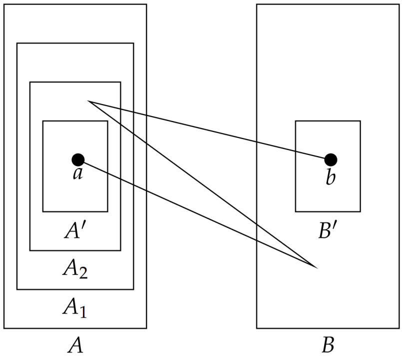
Method Sketch.

Instead of defining X' randomly, define X' by picking $y \in Y$ at random, and let $X' = neighbourhood(y)$. Now determine properties of X', e.g. the expected size of X' is average degree of $y \in Y$. $\square$

```
let ?M = "uniform_count_measure Y"
interpret P: prob_space ?M
  by (simp add: Y_not_empty partitions_finite prob_space_uniform_count_measure)
have sp: "space ?M = Y"
  by (simp add: space_uniform_count_measure)
(* First show that the expectation of the size of X' is the average degree of a vert
have avg_degree: "P.expectation (λ y . card (neighborhood y)) = density * (card X)"
proof -
  have "density = (∑y ∈ Y . degree y)/(card X * card Y)"
    using edge_size_degree_sumY density_simp by simp
  then have d: "density * (card X) = (∑y ∈ Y . degree y)/(card Y)"
    using card_edges_between_set edge_size_degree_sumY partitions_finite(1) partitio
  have "P.expectation (λ y . card (neighborhood y)) = P.expectation (λ y . degree y)
    using alt_deg_neighborhood by simp
  also have "... =(∑ y ∈ Y. degree y)/(card Y)" using P.expectation_uniform_count
    by (simp add: partitions_finite(2))
  finally show ?thesis using d by simp
qed
```

Lemma 3.2

There are many paths of length 3 between vertices in subsets



Lemma 3.2 Diagrammatically (Zhao, 2019)

Lemma 3.2

There are many paths of length 3 between vertices in subsets

Lemma

Let G be a bipartite graph with finite vertex sets X and Y and density δ. Then there are subsets $X' \subseteq X$ and $Y' \subseteq Y$ with $|X'| \geq \delta^2 |X|/16$ and $|Y'| \geq \delta |Y|/4$ such that for every $x \in X'$ and $y \in Y'$ the number of paths of length 3 between x and y in G is at least $\delta^6 |X||Y|/2^{13}$.

```
lemma (in fin_bipartite_graph) walks_of_length_3_subsets_bipartite:
  obtains X' and Y' where "X' ⊆ X" and "Y' ⊆ Y" and
  "card X' ≥ (edge_density X Y)^2 * card X / 16" and
  "card Y' ≥ edge_density X Y * card Y / 4" and
  "∀ x ∈ X'. ∀ y ∈ Y'. card {p. connecting_walk x y p ∧ walk_length p = 3} ≥
  (edge_density X Y)^6 * card X * card Y / 2^13"
```

"Transporting" Information across proofs

Lemma 3.2 involves many different probability spaces ($X2 \subset X$)

```
interpret P1: prob_space "uniform_count_measure X"
interpret P2: prob_space "uniform_count_measure X2"
interpret P3: prob_space "uniform_count_measure Y"
```

... And several graph constructs

```
interpret H: fin_bipartite_graph "(?X1 ∪ Y)" "{e ∈ E. e ⊆ (?X1 ∪ Y)}" "?X1" "Y"
let ?E_loops = "mk_edge ` {(x, x') | x x'. x ∈ X2 ∧ x' ∈ X2 ∧
  (H.codegree_normalized x x' Y) ≥ ?δ ^ 3 / 128}"
interpret Γ: ulgraph "X2" "?E_loops"
```

We can transport information easily using locale definitions

```
have neighborhood_unchanged: "∀ x ∈ ?X1. neighbors_ss x Y = H.neighbors_ss x Y"
  using neighbors_ss_def H.neighbors_ss_def vert_adj_def H.vert_adj_def by auto
then have degree_unchanged: "∀ x ∈ ?X1. degree x = H.degree x"
  using H.degree_neighbors_ssX degree_neighbors_ssX by auto
```

The Graph Theoretic "BSG" Lemma 3.6

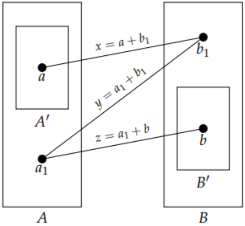
Lemma

Let A be a finite subset of an abelian group G with additive energy $2c$. Then A has subsets B and C with $|B| \geq c^4|A|/16$ and $|C| \geq c^2|A|/4$ such that $|C - B| \leq 2^{13}c^{-15}|A|$.

Proof Elements:

- Create an auxiliary graph using the θ-popular difference: $d \in G$ is θ-popular if $|\{(a, b) \in A^2 | a - b = d\}| \geq \theta|A|$.
- Apply Lemma 3.2.
- Find number of unique sextuples for each $d \in C - B$ based on popularity properties (Lemma 3.5) and paths of length 3.
- Determine bounds on $|C - B|$.

Application of the lemma to additive combinatorics



Credit: Zhao, 2019

Graph Construct

- Vertices: B is a copy of $A \subseteq G$
- Edges: $(x_i, y_i) \in E$ if and only if $y_i - x_i$ is θ-popular.
- Lemma 3.2 then gives large subsets with paths of length 3.

Working in the additive abelian group context:

```
let ?X = "A × {0:: nat}"
let ?Y = "A × {1:: nat}"
let ?E = "mk_edge ` {(x, y)| x y. x ∈ ?X ∧ y ∈ ?Y ∧ (popular_diff (fst y ⊖ fst x) c A)}"
interpret H: fin_bipartite_graph "?X ∪ ?Y" ?E ?X ?Y
```

The Graph Theoretic "BSG" Formalisation

Formalisation Challenges:

- Counting and cardinalities: This statement took 113 lines to prove formally compared to 1 sentence in literature.

```
have card_ineq1: "⋀ x y. x ∈ ?B ⟹ y ∈ ?C ⟹ card ({(z, w) | z w. z ∈ A ∧ w ∈ A ∧
popular_diff (z ⊖ x) c A ∧ popular_diff (z ⊖ w) c A ∧ popular_diff (y ⊖ w) c A}) ≥
(c^12) * ((card A)^2) / 2^13"
```

- Functions for projections: e.g. tuple of 3 pairs from sextuples:

```
define f:: "'a × 'a × 'a × 'a × 'a × 'a ⇒ ('a × 'a) × ('a × 'a) × ('a × 'a)" where
  "f ≡ (λ (p, q, r, s, t, u). ((p, q), (r, s), (t, u)))"
```

Main Theorem On Paper

Theorem (Balog–Szemerédi–Gowers)

Let $c > 0$. For every finite subset of an abelian group A with additive energy $2c$ there exists a subset A' in A of cardinality at least $c^2|A|/4$ such that $|A' - A'| \leq 2^{30}|A|/c^{34}$.

Proof.

1. We simply take $A' = C$ in the previous lemma.
2. and apply the Ruzsa triangle inequality. That tells us that $|B||C - C| \leq |B - C|^2$.
3. which by the lemma is at most $(2^{13}c^{-15}|A|)^2$.
4. Since $|B| \geq c^4|A|/16$ from (1) we obtain the bound stated.

□

Main theorem in Isabelle/HOL

```
theorem Balog_Szemeredi_Gowers: fixes A::"'a set" and c::real
  assumes afin: "finite A" and "A ≠ {}" and "c>0" and "additive_energy A = 2 * c" and ass: "A ⊆ G"
  obtains A' where "A' ⊆ A" and "card A' ≥ c^2 * card A / 4" and
    "card (differenceset A' A') ≤  2^30 * card A / c^34"
proof-
  obtain B and A' where bss: "B ⊆ A" and bne: "B ≠ {}" and bge: "card B ≥ (c^4) * (card A)/16"
    and a2ss: "A' ⊆ A"  and a2ge: "card A' ≥ (c^2) * (card (A))/4"                              (1)
    and hcardle: "card (differenceset A' B) ≤ 2^13 * card A / c^15"
    using assms obtains_subsets_differenceset_card_bound by metis
  have Bg0: "(card B :: real) > 0" using bne afin bss infinite_super by fastforce
  have "(card  B) * card (differenceset A' A') ≤
    card (differenceset A' B) * card (differenceset A' B)"
    using afin a2ss bss infinite_super ass Ruzsa_triangle_ineq1 card_minusset' differenceset_commute   (2)
      sumset_subset_carrier subset_trans sumset_commute by (smt (verit, best))
  then have "card B * card (differenceset A' A') ≤ (card (differenceset A' B))^2"
    using bss bss power2_eq_square by metis
  then have "(card (differenceset A' A')) ≤ (card (differenceset A' B))^2/card B"
    using Bg0 nonzero_mult_div_cancel_left[of "card B" "card(differenceset A' A')"]
      divide_right_mono by (smt (verit) of_nat_0 of_nat_mono real_of_nat_div4)                  (3)
  moreover have "(card (differenceset A' B))^2  ≤ ((2^13) * (1/c^15)*(card A))^2"
    using hcardle  by simp
  ultimately have "(card (differenceset A' A')) ≤ ((2^13) * (1/c^15)*(card A))^2/(card B)"
    using pos_le_divide_eq[OF Bg0] by simp
  moreover have "(c^4) * (card A)/16 >0"
    using assms card_0_eq by fastforce
  moreover have "((2^13) * (1/c^15) * (card A))^2/(card B) =
    ((2^13)* (1/c^15)*(card A))^2 * (1/(card B))" by simp
  moreover have "((2^13)* (1/c^15) * (card A))^2 * (1/(card B)) ≤
    ((2^13) * (1/c^15) * (card A))^2/ ((c^4) * (card A)/16)"                                     (4)
    using bge calculation(2, 3) frac_le less_eq_real_def zero_le_power2 by metis
  ultimately have "(card (differenceset A' A')) ≤ ((2^13) * (1/c^15) * (card A))^2/ ((c^4) * (card A)/16)"
    by linarith
  then have "(card (differenceset A' A')) ≤ (2^30) * (card A)/(c^34)"
    using card_0_eq assms by (simp add: power2_eq_square)
  then show ?thesis using a2ss a2ge that by blast
qed
```

Challenges and Highlights

- The locale approach:
  - Building on previous work - the graph theory library again proved to be modular, easy to work with, and flexible.
  - NEW: Critical to managing the interplay between different mathematical fields.
- Probabilistic Methods:
  - Main challenge = finding theorems, and setting up the formal prob space.
  - Examples of proof structure in combinatorics.
- de Bruijn Factor: 13.9 - cardinality/counting/arithmetic lemmas continue to be key contributors.

Concluding Thoughts

- With 3 contributors, the formalisation took less than 2 months (including new graph theory library).
- Key Contributions: The first formalisation of the Balog–Szemerédi–Gowers Theorem, formal techniques for probabilistic methods in combinatorics, and development of the locale-centric approach.
- Other supplementary results and the sumset alternate also formalised.
- Future work: More additive combinatorics, further development of probabilistic methods.

Acknowledgements and Contacts

Angeliki[1]: ak2110@cam.ac.uk

Mantas[2]: mb2412@cam.ac.uk

Chelsea[3]: cle47@cl.cam.ac.uk

Isabelle AFP Entries:

- `https://www.isa-afp.org/entries/Balog_Szemeredi_Gowers.html`
- `https://www.isa-afp.org/entries/Undirected_Graph_Theory.html`

Funding: This work was funded by the ERC Advanced Grant ALEXANDRIA (Project GA 742178)[1,2], the Cambridge Mathematics Placements (CMP) Internship Programme[2], the Cambridge Trust (Cambridge Australia Scholarships)[3], and a Cambridge Department of Computer Science and Technology Premium Research Studentship[3].

Examples of Additional Results

Lemma

Let A be a finite subset of an Abelian group with $|A + A| \leq C|A|$. Then the additive energy of A is at least C^{-1}.

```
proposition additive_energy_lower_bound_sumset: fixes C::real
  assumes "finite A" and "A ⊆ G" and "(card (sumset A A)) ≤ C * card A" and "card A ≠ 0"
  shows "additive_energy A ≥ 1/C"
```

(And analogous result with difference set).