Parabola Volume 59, Issue 2 (2023)

Isoperimetric sets of English words

Grant Blitz[1], **Timothy Chen**[2], **Marcus Collins**[3], **Milan Duong-Gordley**[4], **Bennett Feng**[5], **William Gao**[6], **Xiaorui Hang**[7], **Summer Kang**[8], **Tina Lou**[9], **Pranav Mallina**[10], **Noah Mok**[11], **Ashwin Naren**[12], **Eve Parrott**[13], **Zubi Talwar**[14], **Aadi Upadhyayula**[15] and **Edward Zhang**[16]

1 Introduction

The isoperimetric problem is among the oldest in mathematics. It asks for the least-perimeter way to enclose a given volume. In this paper, we consider a space of English words and seek subsets of small volumes of minimum or maximum perimeter.

While there has been much study of numerical properties of the English language [4] and of the isoperimetric problem on discrete spaces (see, e.g., [2, 3]), as far as we know, our specific focus is new.

We use a standard list of the 3000 most common English words [1], with a notion of distance and perimeter; see Section 2. We begin with the study of singletons (word sets with volume 1).

Two-letter words

For our space of two-letter words, the singletons of least perimeter are {pc}, {tv}, {hi}, and {up}, with boundaries {pm}, {to}, {he}, and {us}, respectively; see Proposition 4. The singletons of greatest perimeter (7) are {as}, {ie}, and {me}; see Proposition 5.

[1]Grant Blitz is an eighth-grader at Science and Arts Academy in Des Plaines, Illinois.
[2]Timothy Chen is an eighth-grader at Rancho San Joaquin Middle School in Irvine, California.
[3]Marcus Collins is a ninth-grader at the Harvard-Westlake middle school in Los Angeles.
[4]Milan Duong-Gordley is a student at the Proof School in San Francisco.
[5]Bennett Feng is a freshman at Horace Mann School in New York City.
[6]William Gao is student at Canyon Crest Academy in San Diego.
[7]Xiaorui Hang is a student at Lexington High School in Massachusetts.
[8]Summer Kang is a ninth-grader at Proof School in San Francisco.
[9]Tina Lou is a ninth-grader at Vestavia Hills High School in Alabama.
[10]Pranav Mallina is a student at the McCall Middle School in Winchester, Massachusetts.
[11]Noah Mok is a ninth-grader at Davidson Academy Online.
[12]Ashwin Naren is an eighth-grader from Santa Clara, California.
[13]Eve Parrott is a ninth-grader at Livingston High School in New Jersey.
[14]Zubi Talwar is a student at Lakeside Middle School in Seattle.
[15]Aadi Upadhyayula is a student from Southlake, Texas.
[16]Edward Zhang is an eighth-grader at Rancho San Joaquin Middle School in Irvine, California.

Three-letter words

For our space of three-letter words, there are 15 isolated singletons with no boundary; see Proposition 6. They are

$$\{\text{ask}\}, \{\text{ceo}\}, \{\text{dna}\}, \{\text{egg}\}, \{\text{etc}\}, \{\text{eye}\}, \{\text{fly}\}, \{\text{ice}\}, \{\text{its}\}, \{\text{mom}\}, \{\text{mrs}\}, \{\text{off}\}, \{\text{oil}\}, \{\text{via}\}, \{\text{you}\}$$

Note for example that {ice} is isolated only because *ace* is not on the list of 3000 words. The singleton with maximum perimeter (12) is {set}. It has boundary

$$\partial\{\text{set}\} = \{\text{bet, get, jet, let, net, pet, sea, see, sex, sit, wet, yet}\};$$

see Proposition 7.

All words

For the space of all words on our list, there are 1782 singletons with no boundary out of the 3000 words (Proposition 8). The singleton with maximum perimeter (16) is {ear} (Proposition 10). It has boundary

$$\{\text{bar, car, eat, era, far, war, bear, dear, earn, fear, gear, hear, near, tear, wear, year}\}.$$

Doubletons (volume 2)

For the space of two-letter words, the doubleton with minimum perimeter (1) is the pair {pc, pm}, with boundary {am} (Proposition 12). The doubleton with greatest perimeter (14) in the space of two-letter words is $\{\text{as, ie}\}$, which has boundary

$$\partial\{\text{as, ie}\} = \{\text{ad, ah, am, at, be, he, if, in, it, me, ms, us, vs, we}\}$$

(Proposition 13). In the universe of three-letter words, every doubleton with minimum perimeter (0) is formed in one of two ways: the doubleton is made of two isolated singletons or two singletons whose boundaries are each other (Proposition 14). The doubletons with maximum perimeter (21) in the universe of three-letter words are

$$\{\text{set, cap}\}, \{\text{set, gay}\}, \{\text{set, lay}\}, \{\text{set, may}\}$$

(Proposition 15). In the universe of all words, the sets of volume two with minimum perimeter (0) include the $\binom{1782}{2}$ pairs of isolated words (Proposition 16). In the universe of all words, the doubleton with maximum perimeter (31) is {bet, ear} (Proposition 17).

Proofs

Our methods include computation and elementary deductions therefrom. Specifically, computation was used to calculate word sets of given volume and word length and generate tables of word sets and their boundaries and perimeters. The code was written in Python and the results were confirmed with a different Python program implemented independently.

Graph theory

Our study could be phrased in graph theory, as words being the vertices of a graph, and words distance 1 apart connected by a bidirectional edge. Our notion of boundary (Definition 3) is sometimes known as *vertex boundary* in graph theory [3]. We do not utilize graph theory in this paper, but it could be useful for example in finding the distance (the shortest path) between two words. Restrictions on this value could be used to discriminate which words to suggest in spell-checking, as described below. We may also be interested in properties which originate from graphs, such as considering graphs with weighted edges based on frequencies of certain errors.

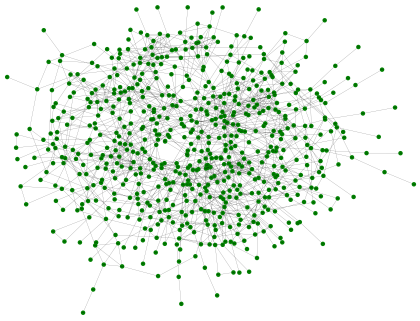

Figure 1 : *The graph of words of at most four letters and at least one neighbour.*

Spell check

Note that a singleton word with large perimeter requires special attention in spell checking, because many single errors could turn it into another legitimate word. Similarly, the appearance of such a word could be due to a single error in a nearby word. We computed that the most problematic 3-letter typo is *het*, at distance 1 from fourteen 3-letter words from our list: bet, get, hat, her, hey, hit, hot, jet, let, net, pet, set, wet, yet. Mac OS and iOS auto-correct just leave it unchanged.

Acknowledgements

This paper is a product of the 2022 MathPath breakout on Metric Spaces, advised by Frank Morgan. The authors thank Dr. Morgan for his advice and guidance throughout the breakout and the MathPath staff for their support. Additionally, we thank Thomas Britz for his suggestion to add graphs that visualize the distance between words.

2 Definitions

We define several colloquial terms for specific use in this paper.

Definition 1 (Space W)**.** *We define the space W as the* 3000 *most common English words as provided by ef.edu [1]. Let W_n denote the set of all n-letter words. The list generally does not include plurals or conjugations — e.g.,* is *is not found in the* 3000 *words. We ignore punctuation and capitalization; for example,* n't*, the contraction for* not*, is considered to be a two-letter English word* nt*,* e-mail *is considered to be a five-letter word* email*, and abbreviations such as* Mr *and* Ms *are considered to be the words* mr *and* ms *respectively.*

Definition 2 (Distance)**.** *Given words w_1 and w_2 in $Z \subset W$, we define the distance from w_1 to w_2 in Z as the minimum number of letter changes it takes to change w_1 into w_2. A letter change is defined as inserting a letter, deleting a letter, replacing a letter, or swapping two adjacent letters. Each word along the path from w_1 to w_2 must be another member of the set Z.*

Definition 3 (Boundary)**.** *The* boundary *in $Z \subset W$ of a set $S \subset Z$ is the set of words in $Z - S$ which are distance one from at least one word in S. We denote the boundary of S as ∂S. The number of elements of the boundary is called the* perimeter*. A word w is* isolated *if $\{w\}$ has an empty boundary.*

3 Two-letter words

Propositions 4 and 5 provide, in the universe W_2 of two-letter words, the sets of volume 1 of minimum and maximum perimeter. They follow from the Table 1 below, which shows several two-letter words and their boundaries, along with their boundary size (perimeter).

Proposition 4. *In the universe W_2 of two-letter words, the singletons of minimum perimeter* (1) *are $\{pc\}$, $\{tv\}$, $\{hi\}$, and $\{up\}$. Their respective boundaries are $\{pm\}$, $\{to\}$, $\{he\}$, and $\{us\}$.*

Proposition 5. *In the universe W_2 of two-letter words, the singletons of maximum perimeter* (7) *are $\{as\}$, $\{ie\}$, and $\{me\}$.*

The boundary of $\{as\}$ is

$$\partial\{as\} = \{ad, ah, am, at, ms, us, vs\}.$$

The boundary of $\{ie\}$ is

$$\partial\{ie\} = \{be, he, if, in, it, me, we\}.$$

The boundary of $\{me\}$ is

$$\partial\{me\} = \{be, he, ie, mr, ms, my, we\}.$$

Proof. Both propositions follow immediately from Table 1. □

Table 1 : *Singletons and their boundaries in the universe W_2 of two-letter words.*

Word Set	Boundary	Boundary Size
{hi}	{he}	1
{pc}	{pm}	1
{tv}	{to}	1
{up}	{us}	1
{by}	{be, my}	2
{pm}	{am, pc}	2
	⋮	
{at}	{ad, ah, am, as, it, nt}	6
{ms}	{as, me, mr, my, us, vs}	6
{no}	{do, go, nt, on, so, to}	6
{on}	{in, no, of, oh, ok, or}	6
{as}	{ad, ah, am, at, ms, us, vs}	7
{ie}	{be, he, if, in, it, me, we}	7
{me}	{be, he, ie, mr, ms, my, we}	7

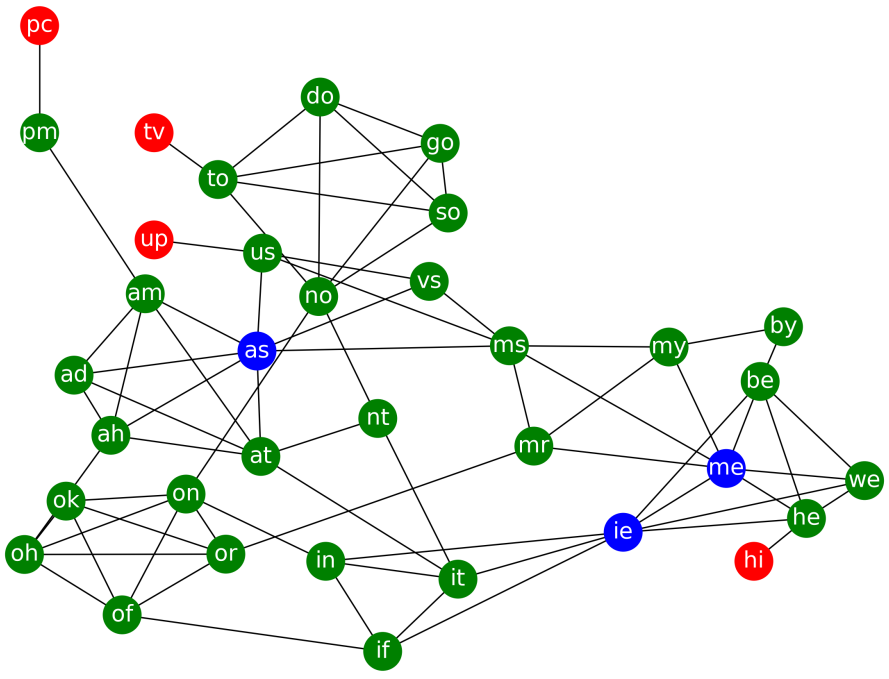

Figure 2 : *The graph of two-letter words, with the four words with just one neighbour in red and the three words with the maximum of seven neighbours in blue.*

4 Three-letter words

Propositions 6 and 7 provide in the universe W_3 of three-letter words the sets of volume 1 of minimum and maximum boundary. They follow from the following Table 2, which shows all three-letter words and their boundaries, along with their boundary size.

Table 2 : *Singletons and their boundaries in the universe W_3 of three-letter words.*

Word Set	Boundary	Boundary Size
{ask}	{}	0
{ceo}	{}	0
{dna}	{}	0
{egg}	{}	0
{etc}	{}	0
{eye}	{}	0
{fly}	{}	0
{ice}	{}	0
{its}	{}	0
{mom}	{}	0
{mrs}	{}	0
{off}	{}	0
{oil}	{}	0
{via}	{}	0
{you}	{}	0
	⋮	
{cap}	{can, car, cat, cop, cup, gap, lap, map, tap}	9
{gay}	{day, gap, gas, guy, lay, may, pay, say, way}	9
{jet}	{bet, get, jew, let, net, pet, set, wet, yet}	9
{lay}	{day, gay, lab, lap, law, may, pay, say, way}	9
{may}	{day, gay, lay, mad, man, map, pay, say, way}	9
{yet}	{bet, get, jet, let, net, pet, set, wet, yes}	9
{let}	{bet, get, jet, leg, lot, net, pet, set, wet, yet}	10
{bet}	{bed, bit, but, get, jet, let, net, pet, set, wet, yet}	11
{net}	{bet, get, jet, let, new, not, nut, pet, set, wet, yet}	11
{pet}	{bet, get, jet, let, net, per, pot, put, set, wet, yet}	11
{set}	{bet, get, jet, let, net, pet, sea, see, sex, sit, wet, yet}	12

Proposition 6. *In the universe W_3 of three-letter words, the singletons of words with minimum perimeter* (0) *are* {*ask*}, {*ceo*}, {*dna*}, {*egg*}, {*etc*}, {*eye*}, {*fly*}, {*ice*}, {*its*}, {*mom*}, {*mrs*}, {*off*}, {*oil*}, {*via*}, *and* {*you*}. *There are* 15 *of these singletons, all of which have empty boundary.*

Proposition 7. *In the universe W_3 of three-letter words, the set of volume 1 and maximum perimeter (12) is $\{set\}$. Its boundary is*

$$\partial\{set\} = \{bet, get, jet, let, net, pet, sea, see, sex, sit, wet, yet\}.$$

Proof. Both propositions follow immediately from Table 2. □

5 All words

Propositions 8 and 10 provide in the universe W of all words the singletons of minimum and maximum boundary. The following propositions follow from a computer search.

Proposition 8. *In the universe W of all words, there are 1782 singletons with empty boundary.*

Proof. This follows directly from a computer search. □

Corollary 9. *For $V \leq 1782$, the minimum perimeter of a set with volume V is 0.*

Proof. For a given V, we take the union of V arbitrary singletons that have perimeter 0. This set will also have perimeter 0 since none of its elements are connected to other words. Thus, we have a set with volume V that has no perimeter. □

Proposition 10. *In the universe W of all words, the subset with volume 1 and maximum perimeter (16) is $\{ear\}$. The boundary of $\{ear\}$ is:*

$$\{bar, car, eat, era, far, war, bear, dear, earn, fear, gear, hear, near, tear, wear, year\}.$$

Proof. This follows directly from a computer search. □

6 Volume two

Earlier sections dealt with singletons, but now we will consider word sets of volume two. We generalize the notion of a boundary for a set of any any volume to be all words that are adjacent to at least one member of the set, but not in the set itself. We will begin exploration with two-letter words. But first, a useful lemma.

Lemma 11. *The perimeter of a doubleton $D = \{w_1, w_2\}$ in any universe is the number of elements in the union of the boundaries of its component words, minus two if the words are adjacent. In other words, we have the equation*

$$\partial\{w_1, w_2\} = |\partial\{w_1\} \cup \partial\{w_2\}| - \begin{cases} 2, & \textit{if the singletons are adjacent} \\ 0, & \textit{otherwise} \end{cases}$$

Proof. If a word is in the boundary of D, it is within distance 1 of a word in D, and consequently within the union of the boundaries of those word singletons. However, we choose not to include the words themselves as part of the boundary, hence the subtraction. On the other hand, if a word is in that union, it is within distance 1 of w_1 or w_2, so if in addition it is not in D, it is in the boundary of D. □

Proposition 12. *In the universe W_2 of two-letter words, the set of volume 2 of minimum perimeter* (1) *is* {*pm, pc*}*, with boundary* {*am*}*.*

Proof. Note from Table 1 that for every singleton in W_2, the perimeter is at least 1. Consider a set of volume 2 $\{w_1, w_2\}$ with perimeter at most 1.

Neither of the singletons $\{w_1\}$ or $\{w_2\}$ can have perimeter greater than 2, or by Lemma 11, $\{w_1, w_2\}$ would have perimeter greater than 1. If both singletons had perimeter 1, by Table 1, the perimeter would be 2 because the boundaries of all singletons with perimeter 1 are disjoint. Therefore, w_1 is *by* or *pm* since they are the only singletons with perimeter 2.

If w_1 is *by*, then w_2 must be *my* or *be* because they are the only words in the boundary of {by}. But w_2 cannot be *my* or *be*, because those singletons have perimeter greater than 2. Therefore, w_1 must be *pm*. This means that w_2 must be *am* or *pc*, the words in the boundary of $\{pm\}$. However, $\{am\}$ has boundary greater than 2, so w_2 must be *pc*. The boundary of $\{pm, pc\}$ is $\{am\}$, so it is the unique doubleton with minimum perimeter 1. □

Proposition 13. *In the universe W_2 of two-letter words, the set of volume 2 and maximum perimeter* (14) *is* {*as, ie*}*. The boundary is*

$$\partial\{as, ie\} = \{ad, ah, am, at, be, he, if, in, it, me, ms, us, vs, we\}.$$

Proof. Suppose the maximum perimeter of $\{w_1, w_2\}$ is 14 or more. By Lemma 11 and Table 1, each singleton would need perimeter 7 and no overlap with the other singleton. Only $\{as, ie\}$ satisfies these conditions. □

The preceding two propositions may also be verified by computation, as summarized in Table 3.

Table 3 : *Doubletons and their boundaries in the universe of two-letter words.*

Word Set	Boundary	Boundary Size
{pc, pm}	{am}	1
{hi, pc}	{he, pm}	2
{hi, tv}	{he, to}	2
{hi, up}	{he, us}	2
{pc, tv}	{pm, to}	2
{pc, up}	{pm, us}	2
{tv, up}	{to, us}	2
	⋮	
{as, me}	{ad, ah, am, at, be, he, ie, mr, ms, my, us, vs, we}	13
{as, no}	{ad, ah, am, at, do, go, ms, nt, on, so, to, us, vs}	13
{as, on}	{ad, ah, am, at, in, ms, no, of, oh, ok, or, us, vs}	13
{at, me}	{ad, ah, am, as, be, he, ie, it, mr, ms, my, nt, we}	13
{ie, no}	{be, do, go, he, if, in, it, me, nt, on, so, to, we}	13
{me, no}	{be, do, go, he, ie, mr, ms, my, nt, on, so, to, we}	13
{me, on}	{be, he, ie, in, mr, ms, my, no, of, oh, ok, or, we}	13
{as, ie}	{ad, ah, am, at, be, he, if, in, it, me, ms, us, vs, we}	14

Proposition 14. *In the universe W_3, the 107 sets of volume 2 with minimum perimeter (0) are formed in one of two ways. 105 of the sets come from two isolated singletons. The other two come from two singletons whose boundaries are each other: $\{age, ago\}$ and $\{all, ill\}$.*

Proof. By Lemma 11, if a doubleton has no boundary, either each singleton subset has no boundary or each singleton subset is the boundary of the other. (Note that if one is contained in the boundary of the second, then the second is contained in the boundary of the first). Since by Proposition 6 there are 15 isolated words, there are $\binom{15}{2} = 105$ pairs of them. By computation, there are only two pairs of the second type: $\{age, ago\}$ and $\{all, ill\}$. □

Proposition 15. *In the universe W_3 of three-letter words, the sets of volume 2 and maximum perimeter (21) are $\{set, cap\}, \{set, gay\}, \{set, lay\}$ and $\{set, may\}$.*

Proof. Consider a doubleton $\{w_1, w_2\}$ with perimeter greater than or equal to 21. By Lemma 11, the only possibilities for the two singletons $\{w_1\}, \{w_2\}$ are the 11 of Table 2 with perimeter at least 9 (and at most 12 in the case of $\{set\}$). If both w_1 and w_2 end in *et*, then the boundaries of the singletons have too much overlap for the perimeter of $\{w_1, w_2\}$ to reach 21. Therefore one of them must be *cap*, *gay*, *lay*, or *may*, all with perimeter 9, and the other must be *set* with perimeter 12. The resulting four doubletons indeed have perimeter 21 and are therefore the only maxima of perimeter. □

Proposition 16. *In the universe W of all words, the sets of volume 2 of minimum perimeter (0) include the $\binom{1782}{2} = 1,586,871$ pairs of isolated words.*

Proof. By computation, there are 1782 isolated words. For each of the $1,586,871$ pairs of these words, their set will have perimeter 0. □

Proposition 17. *In the universe W of all words, the set of volume 2 and maximum perimeter (31) is $\{bet, ear\}$. The boundary is {be, bed, bit, but, get, jet, let, net, pet, set, wet, yet, beat, belt, best, bar, car, eat,era, far, war, bear, dear, earn, fear, gear, hear, near, tear, wear, year}.*

Proof. By computation, the two singletons with largest perimeter are {bet} and {ear}, with perimeters of 15 and 16 respectively. By computation, and as referenced in Proposition 10, the boundary of the singleton {ear} is

{bar, car, eat, era, far, war, bear, dear, earn, fear, gear, hear, near, tear, wear, year}.

By computation, the boundary of the singleton {bet} is

{be, bed, bit, but, get, jet, let, net, pet, set, wet, yet, beat, belt, best}.

Since the two boundaries and {bet, ear} are pairwise disjoint, by Lemma 11 the boundary of the doubleton $\{bet, ear\}$ is the union of the boundaries of the singletons, so the boundary of $\{bet, ear\}$ is as asserted, with the maximal 31 elements. □

7 Open questions

(a) A *connected component* is a set with empty boundary and no proper subsets with empty boundary. How many connected components are there in W? There are at least 1782, as there are 1782 singleton components.

(b) Which connected component of W has largest volume? It likely contains the words a and i.

(c) What is the distribution of the perimeter of word sets with given volume?

(d) What is the largest clique in W (where every word is unit distance from every other)?

(e) What sets have minimum and maximum fattened boundary (including words within distance two)?

References

[1] ef.edu, 3000 most common words in English, 2022, https://www.ef.edu/english-resources/english-vocabulary/top-3000-words/, last accessed on 2023-08-03.

[2] S.G. Bobkov and F. Götze, Discrete isoperimetric and Poincare-type inequalities, *Probability Theory and Related Fields* **114** (1997), 245–277, https://www-users.cse.umn.edu/~bobko001/papers/1999_PTRF_BG.pdf, last accessed on 2023-08-03.

[3] F. Chung, Discrete isoperimetric inequalities, *Surveys in Differential Geometry* **9** (2004), 53–82, https://mathweb.ucsd.edu/~fan/wp/iso.pdf, last accessed on 2023-08-03.

[4] Wikipedia, Most common words in English, https://en.wikipedia.org/wiki/Most_common_words_in_English, last accessed on 2023-08-03.