The largest vocabulary without a crossword

Kenny Peng[*]

Abstract

We show that for an alphabet of size n, the maximum vocabulary of two-letter words without a 2×2 crossword has size $(\frac{1}{4} + o(1))n^2$.

1 Introduction

Given n letters, there are n^k possible k-letter words. A vocabulary is a subset of these words. For a given vocabulary, a $k \times k$ crossword is a $k \times k$ grid of letters such that each row and column (read left-to-right and top-to-bottom) forms a distinct word in the vocabulary. For example, the following 5×5 crossword appeared in the New York Times:

m	e	r	c	h
a	g	i	l	e
n	a	v	e	l
i	d	e	a	l
a	s	t	r	o

It is not clear that one can construct a crossword given a vocabulary. It has been shown, for example, that filling in a crossword is computationally hard (e.g., Engel et al. (2012), Gourvès et al. (2024)).

Here, we consider when the existence of a crossword is guaranteed. Let $C(n, k)$ be—given n letters to choose from—the largest number of k-letter words in a vocabulary without a $k \times k$ crossword. We focus only on the case $k = 2$, proving the following result.

Theorem 1. $C(n, 2) = \left(\frac{1}{4} + o(1)\right) n^2$.

Noting that the total number of possible two-letter words is n^2, Theorem 1 implies that given somewhat more than a $1/4$ fraction of possible two-letter words, it is always possible to construct a 2×2 crossword—and that there exists $n^2/4$ words where it is not. Analyzing $C(n, k)$ for $n \geq 3$ is a natural further question.

As we will see, the upper bound in Theorem 1 is implied by considering the maximum number of edges in a digraph with the property that there is at most one path of length two connecting any vertex u to a different vertex v. This is reminiscent of finding the maximum number of edges in an undirected graph without a 4-cycle, which Reiman (1958) showed to be essentially $\frac{n}{4}(1 + \sqrt{4n - 3})$. We employ a similar double-counting approach.

[*] Cornell University, Department of Computer Science (kennypeng@cs.cornell.edu). The question considered here arose through a conversation with Gabriel Agostini and Danice Ball.

2 Proof

Proof of Theorem 1. The lower bound follows from a simple example: Separate n letters into two equal parts A_1 and A_2. Then consider the $n^2/4$ words with first letter in A_1 and second letter in A_2. This vocabulary is crossword free. Indeed, if there existed some crossword

$$\begin{array}{|c|c|}\hline a & b \\ \hline c & d \\ \hline \end{array} \tag{1}$$

then ab and bd would both be words in the vocabulary, implying that b is in both A_1 and A_2, giving a contradiction. (It is not hard to see that one can add additional words to this example while maintaining the crossword-free property. Theorem 1 implies that one cannot, however, add significantly more.)

We now show the upper bound. We do this by showing that any digraph with sufficiently many edges must contain four distinct vertices a, b, c, d such that $a \to b \to d$ and $a \to c \to d$. To see why this suffices, consider a digraph with n vertices, where each vertex corresponds to a letter. Draw an edge from u to v if and only if uv is a word in the vocabulary. Then the existence of

$$\begin{tikzcd} a \arrow[r] \arrow[d] & b \arrow[d] \\ c \arrow[r] & d \end{tikzcd} \tag{2}$$

in the digraph implies the existence of the crossword shown in (1). In what follows, we ignore words with the same letter, which would correspond to self loops. There are at most $n = o(n^2)$ such words, so they can be safely ignored. Therefore, we only consider digraphs without self loops (and without parallel edges).

It suffices to show that for all $\epsilon > 0$, for n sufficiently large, any digraph with n vertices and at least $(\frac{1}{4} + \epsilon)n^2$ edges must contain an occurrence of (2). Let vertex v have indegree $d^-(v)$ and outdegree $d^+(v)$. We count in two ways the number of triples (u, v, w) of vertices (not necessarily distinct) such that $u \to v \to w$. The number of such triples is equal to

$$\sum_v d^-(v) \cdot d^+(v), \tag{3}$$

where we count the number of times a vertex v appears in the middle position of a triple. Furthermore, each pair of vertices (u, w) can only be part of one such triple (u, v, w) when $u \neq w$, since otherwise, there would be a crossword. Therefore, there are at most $n(n-1)$ valid triples of the form (u, v, w) where $u \neq w$. There are also at most $n(n-1)$ valid triples of the form (u, v, u). So in total, we have that

$$\sum_v d^-(v) \cdot d^+(v) \leq 2n(n-1). \tag{4}$$

It now suffices to show that for all $\epsilon > 0$, if there are more than $(\frac{1}{4} + \epsilon)n^2$ edges, then for n sufficiently large, $\sum_v d^-(v) \cdot d^+(v)$ must exceed $2n(n-1)$. Accordingly, we search for a set of vertices with both large indegree and outdegree.

Consider a digraph with n vertices and at least $(\frac{1}{4} + \epsilon)n^2$ edges. Let S be the set of vertices with outdegree at least $\frac{\epsilon}{3}n$. Then

$$\sum_{v \in S} d^-(v) \cdot d^+(v) \geq \frac{\epsilon}{3} n \cdot \sum_{v \in S} d^-(v). \tag{5}$$

Now suppose that there are αn vertices in S and βn vertices in S with outdegree at least $(1-\alpha+\frac{\epsilon}{3})n$. Since there are only $(1-\alpha)n$ vertices outside of S, each vertex with outdegree at least $(1-\alpha+\frac{\epsilon}{3})n$ must have at least $\frac{\epsilon}{3}n$ outgoing edges that point back to vertices in S. Therefore,

$$\sum_{v\in S} d^-(v) \geq \beta n \cdot \frac{\epsilon}{3}n. \tag{6}$$

We will show shortly that $\beta > \frac{\epsilon}{3}$. Assuming this fact for now, it follows from (5) and (6) that

$$\sum_{v\in S} d^-(v)\cdot d^+(v) \geq \frac{\epsilon}{3}n \cdot \frac{\epsilon}{3}n \cdot \frac{\epsilon}{3}n = \frac{\epsilon^3 n^3}{27}. \tag{7}$$

For all n sufficiently large, this is greater than $2n(n-1)$, giving the desired conclusion. It remains to prove the bound on β.

Proof that $\beta > \frac{\epsilon}{3}$. Since all vertices not in S have outdegree at most $\frac{\epsilon}{3}n$,

$$\sum_{v\notin S} d^+(v) < \sum_{v\notin S} \frac{\epsilon}{3}n \leq \frac{\epsilon}{3}n^2. \tag{8}$$

Therefore, since $\sum_v d^+(v) \geq (\frac{1}{4}+\epsilon)n^2$,

$$\sum_{v\in S} d^+(v) \geq \left(\frac{1}{4}+\epsilon\right)n^2 - \sum_{v\notin S} d^+(v) > \left(\frac{1}{4}+\frac{2\epsilon}{3}\right)n^2. \tag{9}$$

Now recall that there are αn vertices in S and βn vertices in S with outdegree at least $(1-\alpha+\frac{\epsilon}{3})n$. Then

$$\sum_{v\in S} d^+(v) < \beta n \cdot n + (\alpha n - \beta n)\cdot\left(1-\alpha+\frac{\epsilon}{3}\right)n, \tag{10}$$

where we use that $d^+(v) < n$ for all v. Simplifying,

$$\begin{aligned}\sum_{v\in S} d^+(v) &< n^2\left(\beta + (\alpha-\beta)\left(1-\alpha+\frac{\epsilon}{3}\right)\right) & (11)\\ &= n^2\left(\beta+\alpha-\alpha^2+\frac{\epsilon}{3}\alpha-\beta+\alpha\beta-\frac{\epsilon}{3}\beta\right) & (12)\\ &= n^2\left(\alpha(1-\alpha)+\alpha\beta+\frac{\epsilon}{3}(\alpha-\beta)\right) & (13)\\ &\leq n^2\left(\alpha(1-\alpha)+\beta+\frac{\epsilon}{3}\right), & (14)\end{aligned}$$

where in the last line we use that $\alpha, \alpha-\beta \leq 1$. Noticing that $\alpha(1-\alpha) \leq \frac{1}{4}$ for all α,

$$\sum_{v\in S} d^+(v) < n^2\left(\frac{1}{4}+\beta+\frac{\epsilon}{3}\right). \tag{15}$$

Combining with (9), we have that

$$\left(\frac{1}{4}+\beta+\frac{\epsilon}{3}\right)n^2 > \left(\frac{1}{4}+\frac{2\epsilon}{3}\right)n^2, \tag{16}$$

which implies that $\beta > \frac{\epsilon}{3}$, as desired.

References

J. Engel, M. Holzer, O. Ruepp, and F. Sehnke. On computer integrated rationalized crossword puzzle manufacturing. In *Fun with Algorithms: 6th International Conference, FUN 2012, Venice, Italy, June 4-6, 2012. Proceedings 6*, pages 131–141. Springer, 2012.

L. Gourvès, A. Harutyunyan, M. Lampis, and N. Melissinos. Filling crosswords is very hard. *Theoretical Computer Science*, 982:114275, 2024.

I. Reiman. Über ein problem von K. Zarankiewicz. *Acta Mathematica Academiae Scientiarum Hungarica*, 9:269–273, 1958.