SLAC – PUB – 4085
September 1986
M

DYNAMIC STUDY OF THE GRAPHS KNN AND KMST*

VITO DI GESU'

Dip.di Matematica e Applicazioni,
Univ.di Palermo, ITALY
and
Stanford Linear Accelerator Center
Stanford University, Stanford, California, 94305

ABSTRACT

This paper deals with the dynamics of the growing of two subgraphs, the K-Nearest Neighbour (KNN) and the K-Minimum Spanning Tree ($KMST$) of an undirected weighted and complete graph G. Both are widely used in cluster and data analysis. Some limit values of K (K_{max}), for which they become complete, are given. The results of some experiments on "random graphs" are also presented.

Key words: *graph theory, KMST, KNN, Convex Hull.*

Submitted to *Discrete Mathematics*

⋆ Work supported by the Department of Energy, contract DE – AC03 – 76SF00515.

1. INTRODUCTION

The dynamic study of a "system" provides useful information about its finer nature (presence of layered configurations) and allows one to discover changes in its space-states. Dynamic study of graphs is of great interest in this context. Given a weighted graph G, the computation of the number of components of its minimum spanning forest for increasing thresholds [1], and its K-nearest neighbours for increasing K [2] are examples of dynamic analysis.

In this paper we analyse some dynamic properties of two subgraphs: the K-nearest neighbour (KNN) and K-Minimum Spanning Tree ($KMST$) of an undirected weigthed and complete graph G [3,4,5]. All those features, that depend from the variable K, are considered as dynamic. Some topological configurations, that give some information about the distribution of the nodes in the case of random graph, are studied. The results obtained may be useful for the analysis of "sparse images" [6] and the study of multidimentional clustering problems [3].

The next section is dedicated to some definitions and notation used througout the paper. In Section Three some limit values of K, for which the $KMST$ and the KNN become complete, are derived. A relation between these two subgraphs is also established. In Section Four two topological configurations of nodes ("Pure Star" and "Pure Linear") are investigated. In Section Five some experiments are presented, concerning the dynamic behavour in the case of "random graphs". Section Six presents some concluding remarks.

2. DEFINITIONS AND NOTATION

In the following undirected, complete, and weigthed graphs $G =< N, W >$, are considered. Here N is the set of nodes, the cardinality of which is denoted by $|N|$, $W: N \times N \rightarrow R^+$ is a norm function and R^+ is the set of positive real numbers. The nodes of G are elements of a d-dimentional space $\mathbf{X}$.

For each node $x \in N$ the set $N_x = N - \{x\}$ is ordered as follows:

$$\forall y, z \in N_x \quad y \preceq z \iff W(x,y) \leq W(x,z)$$

The first K nodes of N_x are the K-nearest neighbours linked to x.

DEF 1. The degree of a node $x \in N$, OD_x, is the number of arcs linked to it.

DEF 2. The $KNN(G)$ is a subgraph of G,such that each node $x \in N$ is linked with its K nearest nodes.

DEF 3. The $MST(G)$ is a spanning tree of G, such that the sum of the weigth of its arcs is minimal.

DEF 4. The $KMST(G)$ is a spanning subgraph of G, such that:

$$1MST(G) = MST(G)$$

$$KMST(G) = [K-1]MST(G) \bigcup MST(G - [K-1]MST(G)) \ K > 1.$$

Here "$G - [K-1]MST(G)$" denotes the set difference between graphs.

In the following $G - KMST(G) \equiv G^{(K)} =< N^{(K)}, W^{(K)} >$. Here $N^{(K)} = N - \{x | \ x \in N \ and \ OD_x = |N| - 1\}$, and $W^{(K)} = W - \{W | \ W \in [K-1]MST(G)\}$. Figure.1 displays the nodes of G, Figure.2a and Figure.2b show the 2NN and the 2MST of G respectively.

DEF 5. A graph such that $\exists e_1, e_{|N|} \in N, OD_{e_1} = OD_{e_{|N|}} = 1$ and $\forall e_i \in N - \{e_1, e_{|N|}\}, OD_{e_i} = 2$ is a chain and will be denoted by $(e_1, e_2, ..., e_i, ..., e_{|N|})$.

3. DYNAMICS OF THE *KNN* AND *KMST*

This section presents some results concerning the value of K for which the $KNN(G)$ and the $KMST(G)$ become equal to G. This value is denoted by K_{max}. An exact value of K_{max} is found for the $KNN(G)$, while for the $KMST(G)$ only some limit values are extabilished.

If the nodes of G are points in a normed space $\mathbf{X}$, the following lemma holds:

LEMMA 1. Let T be the subgraph added to the $KNN(G)$ to compute the $[K+1]NN(G)$, then T is a forest.

Suppose that a cycle of order L exists:

$$(x_1, x_2, \dots, x_{L-1}, x_1)$$

It is easy to see that the following order may be stated between the weigths of the arcs in the cycle:

for $i = 1, 2, ..., L-3$

x_{i+1} *is linked to* $x_i \iff W(x_i, x_{i+1}) \leq W(x_{i+1}, x_{i+2});$

x_{L-1} *is linked to* $x_{L-2} \iff W(x_{L-2}, x_{L-1}) \leq W(x_{L-1}, x_1);$

and finally:

x_1 *is linked to* $x_{L-1} \iff W(x_{L-1}, x_1) < W(x_1, x_2);$

The last inequality is strict because x_1 has yet to be linked to x_2 at the beginning of the cycle and this leads to the contradiction $W(x_1, x_2) < W(x_1, x_2)$. Therefore T is a forest $\triangle$.

LEMMA 2. For each $1 \leq K < |N| - 1$, $\exists x \in N$ of the $KNN(G)$ such that $OD_x = K$.

For $K = 1$ this follows from *LEMMA 1* (1NN is a forest). Suppose it is true for $1 < K \leq |N| - 1$, then is true for $K + 1$. In fact let $Y = \{y \mid OD_y = K\}$, then to compute the $[K+1]NN(G)$ only $|Y| - 1$ edges are linked to the nodes of Y and by *LEMMA 1* they make a forest. Therefore $\exists y \in Y$ such that $OD_y = K + 1$ $\triangle$.

THEOREM 1. For the $KNN(G)$ the value of K_{max} is $|N| - 1$.

It follows from *LEMMA 2* $\triangle$.

Note that the result is valid only if the space **X** is normed.

The evaluation of the K_{max} of the $KMST(G)$ is more difficult and depends strongly on the configuration of the nodes in **X**. Two limit conditions will be stated below. Two very uncommon configurations of the nodes are considered for this purpose.

DEF 6. A graph G is said pure linear, PL, if its nodes are topologically configured as follows:

$$\begin{aligned}
&\forall 0 \leq K < |N| \;\; \exists(e_1, e_2, ..., e_{|N^{(K)}|}) \Longrightarrow \\
&[K+1]MST(G) = KMST(G) \bigcup (e_1, e_2, ..., e_i ..., e_{|N^{(K)}|})
\end{aligned}$$

Figure.3 shows an example of a PL graph.

DEF 7. A graph G is said pure star, PS, if its nodes are topologically configured as follows:

$$\begin{aligned}
&\forall 0 \leq K < |N| \;\; \exists x_K \in N^{(K)} \Longrightarrow \\
&W(x_K, y) \leq min\{W(y,z) | y, z \in N^{(K)}\}
\end{aligned}$$

The node x_K is called the dominant of $G^{(K)}$ (x_K **dom** $G^{(K)}$). Figure.4 gives an example of a PS graph.

The two classes of graphs are denoted by **PL** and **PS** respectively.

LEMMA 3. The value of K_{max} for $G \in \mathbf{PL}$ is $\lceil |N|/2 \rceil$ and this is the minimum value.

In fact, by *DEF 6*, at each step of the computation of the $KMST(G)$ there are two cases:

$$\begin{aligned}
&|N| \; even \Longrightarrow \\
&\quad \forall 0 \leq K < |N| \;\; |N^{(K)}| = |N| \; and \\
&\quad K_{max} \times (|N| - 1) = |N| \times (|N| - 1)/2 \Longrightarrow K_{max} = |N|/2 \\
&|N| \; odd \Longrightarrow \\
&\quad \forall 0 \;\leq\; K < (|N| - 1)/2 \;\; |N^{(K)}| = |N|, \\
&\quad K = (|N| - 1)/2 \;\; |N^{(K)}| = (|N| - 1)/2 \; and \\
&\quad (K_{max} - 1) \times (|N| - 1) + (|N| - 1)/2 = |N| \times (|N| - 1)/2 \Longrightarrow \\
&\quad K_{max} = (|N| + 1)/2
\end{aligned}$$

The value is the minimum because at each step we add the maximum number of edges $\triangle$.

LEMMA 4. The value of K_{max} for $G \in \mathbf{PS}$ is $|N|-1$ and this is the maximum value.

By *DEF 7*, at each step of the computation of the $[i]MST(G)$ $|N|-i$ edges are added and therefore:

$$\sum_{i=1}^{K_{max}} (|N|-i) = |N| \times (|N|-1)/2 \Longrightarrow K_{max} = |N|-1$$

This value is the largest possible because the maximum degree of the node in G is $|N|-1$ $\triangle$.

From the *LEMMAs* 3 and 4 it follows:

THEOREM 2. For a generic G:

$$\lceil |N|/2 \rceil \leq K_{max} \leq |N|-1$$

$\triangle$.

Figure.5 shows an example of graph with $\lceil |N|/2 \rceil < K_{max} < |N|-1$.

From *LEMMA 4* and *DEF. 7* follows that if $G \in \mathbf{PS}$ then at the end of the computation of the $[N-1]MST(G)$ the nodes are ordered in a chain:

$$(x_1, x_2, ..., x_{|N|})$$

$$x_i \prec x_j \iff i < j \; and \; x_i \text{ dom } G^{(i)},\; x_j \text{ dom } G^{(j)}$$

Another dynamic property relating the $KNN(G)$ to the $KMST(G)$ may be stated from the previous results.

THEOREM 3. For each K the $KNN(G)$ is a subgraph of the $KMST(G)$.

The proof is made by induction.

$K = 1 \Longrightarrow 1NN \subseteq 1MST$ true by definition of MST;

suppose it is true that

$K > 1 \Longrightarrow KNN \subseteq KMST$;

then by *LEMMA 1* the subgraph T added to $KNN(G)$ to compute the $[K+1]NN(G)$ is a forest and therefore $T \bigcup KNN(G) \subseteq [K+1]MST(G)$ $\triangle$.

4. CHARACTERIZATION OF THE GRAPHS PS AND PL

In this section we give some conditions on the cardinalty of graphs **PS** and **PL**, whenever their nodes are points of a multidimentional normed space **X**. In the following the dimention of **X** will be denoted by **d**.

Let us first analyse the case **d** = 1. For sake of semplicity the nodes of G are considered ordered from left to rigth in a chain $(1, 2, ..., n-1, n)$.

It is easy to show, by construction, that the class **PS** contains the chains $\{(1), (1,2), (1,2,3)\}$. For $n \geq 4$ the chain $(1, 2, ..., n)$ contains the kernel $(2, 3, ..., n-1)$, in fact the node "2" is the closest to "1" and the node "$n-1$" is the closest to "n" therefore $G \notin$ **PS**.

The following *LEMMA* holds, for the class **PL**:

LEMMA 5. If **d** = 1 then the class **PL** contains only the chains:

$$\{(1), (1,2), ..., (1,2,3,4,5)\}$$

The proof is made by construction. The hypothesis of ordering allows one to compute a K spanning tree of G, $KST(G)$ as follows:

```
$\forall\ 1 \leq K < n$
 begin
  $N' \leftarrow \phi$
  $insert \leftarrow false$
  $j \leftarrow 1$
  while $(|N'| < n)$ do
   begin
    $S \leftarrow 1$
    while $(S < n\ \wedge\ \neg insert)$ do
     begin
      $l = (j + S) \bmod n$
      if $(j,l) \notin KST(G)\ \wedge\ (j,l)$ does not form a cycle then
        $KST(G) \leftarrow KST(G) \cup (j,l)$
        $N' \leftarrow N' \cup (j,l)$
        $j \leftarrow l$
        $insert \leftarrow true$
      else
        $S \leftarrow S + 1$
        $insert \leftarrow false$
      endif
     end
   end
 end.
```

In order to demonstrate the correcteness of this algorithm note that at each step K the inserted arcs do not form cycle. Two cases must be considered:

$K < \lceil n/2 \rceil \Longrightarrow$ the computed tree is a chain of cardinality n. In fact:

$$\forall\, r \neq j \leq N \quad (j + K) \bmod n \neq (r + K) \bmod n$$

If n is even it is easy to verify that after $n - 2/2$ steps the $KST(G)$ is completed in one more step by inserting $n - 1$ edges. Moreover $n/2$

edges must be linked to the node "1".

If n is odd it is easy to verify that after $(n-1)/2$ steps the $KST(G)$ is completed in one more step by inserting $(n-1)/2$ edges. Moreover $(n-1)/2$ edges must be linked to the node "1". The $KMST(G)$ is also a $KST(G)$.

Therefore $G \in \mathbf{PL}$ only if $n \leq 5$. The proof is completed by checking, with the previous algorithm that $\{(1),(1,2),...,(1,2,3,4,5)\} \equiv \mathbf{PL}$ $\triangle$.

COROLLARY If $\mathbf{d} = 1 \Longrightarrow \forall\, G \;\; K_{max} \leq \lceil |N| \rceil /2$.

For $\mathbf{d} > 1$ an upper limit for the cardinality of N is difficult to determine for both classes.

Below, we give some conditions on the spacial distribution of the nodes in $\mathbf{X}$ for the class $\mathbf{PS}$.

The concept of convex hull of a set of points plays a central role in many problems related to pattern recognition, classification [7,8,9] and computational geometry [10]. Relevant information, regarding the distribution of the nodes for $G \in \mathbf{PS}$, may be carried out by considering the convex hull of the set N.

DEF 8. Given a set of points $N \subseteq \mathbf{X}$, the convex hull of N, $CH(N)$, is the smallest polyhedron containing N.

The hypervolume surrounded by the $CH(N)$ is denoted by $\mathbf{int}(CH(N))$. The sum of the edges of the $CH(N)$ is denoted by $\mathbf{Perim}(CH(N))$

THEOREM 4. If $G \in \mathbf{PS}$ then the $\mathbf{int}(CH(N))$ may contains only one point.

Assume that $G \in \mathbf{PS}$ and that $\exists a, b \in \mathbf{int}(CH(N))$. One of the two nodes, suppose a, must be **dom** G. In fact from the triangular inequality it follows that:

$$\sum_{y \in CH(N)} W(a,y) \leq \mathbf{Perim}(CH(N)/d)$$

$$and\ \forall\ y \in CH(N)$$

$$\sum_{z \in CH(N)} W(y,z) > \mathbf{Perim}(CH(N))$$

Consider now a partition of $\mathbf{X}$ in to a set of d orthogonal hyperplanes, such that one of them contains $\{a,b\}$ and then have their origin at the middle of the segment $[a,b]$. By hypothesis $b \in \mathbf{int}CH(N)$. Then $\exists z \in CH(N)$ such that $\{b,z\}$ is in the same partition of $\mathbf{X}$, therefore:

$$W(b,z) \ \leq \ W(a,z)$$

The last inequality contraddicts that hypothesis that a is **dom** G. Therefore only one point may belong to $\mathbf{int}(CH(N))\triangle$.

The last result allows one to state that for $\mathbf{d} = 2$, if $G \in \mathbf{PS}$ then $|N| \leq 5$.

Computer experiments, on simulated "random graphs", made for $\mathbf{d} > 2$ seem to support the following conjecture: may be claimed:

$$G \in \mathbf{PS} \Longrightarrow |N| \leq 2 \times \mathbf{d} + 1$$

5. BEHAVIOUR IN THE CASE OF RANDOM GRAPH

Some Monte Carlo simulations have been made in order to study the growing of the $KMST$ of a "random graph". Here a graph is considered random if its nodes are randomly distributed from a uniform distribution in $\mathbf{X}$.

Figure.6 shows the experimental results of K_{max} versus $|N|$ for a random graph. The experimental points fit very well to a straigth line with slope $m = .58$. This value is roughly the average between the two limit conditions given by *THEOREM 2*. The values of K_{max} seem not be related to the dimention of **X**.

A second experiment was made in order to give an estimate of the upper bound for the cardinality of the classes **PL** and **PS**. A sample of 1200 graphs were generated for $d = 1, 2, 4, 6, 8, 10$ and $|N| = 5, 7, 10, 20$. The results are shown in Table 1.

Columns 3 and 4 give the percentage of graphs **PS** and **PL** found in the sample, column 5 indicates the mean value of K_{max}. The experiment seems **to show that the class PL is larger then the class PS, in spite of the fact that** its characterization is more complex.

The data contained in the Table 1, also, confirm the results stated in *THEOREMs 2* and *4* and in *LEMMA 5*.

6. FINAL REMARKS

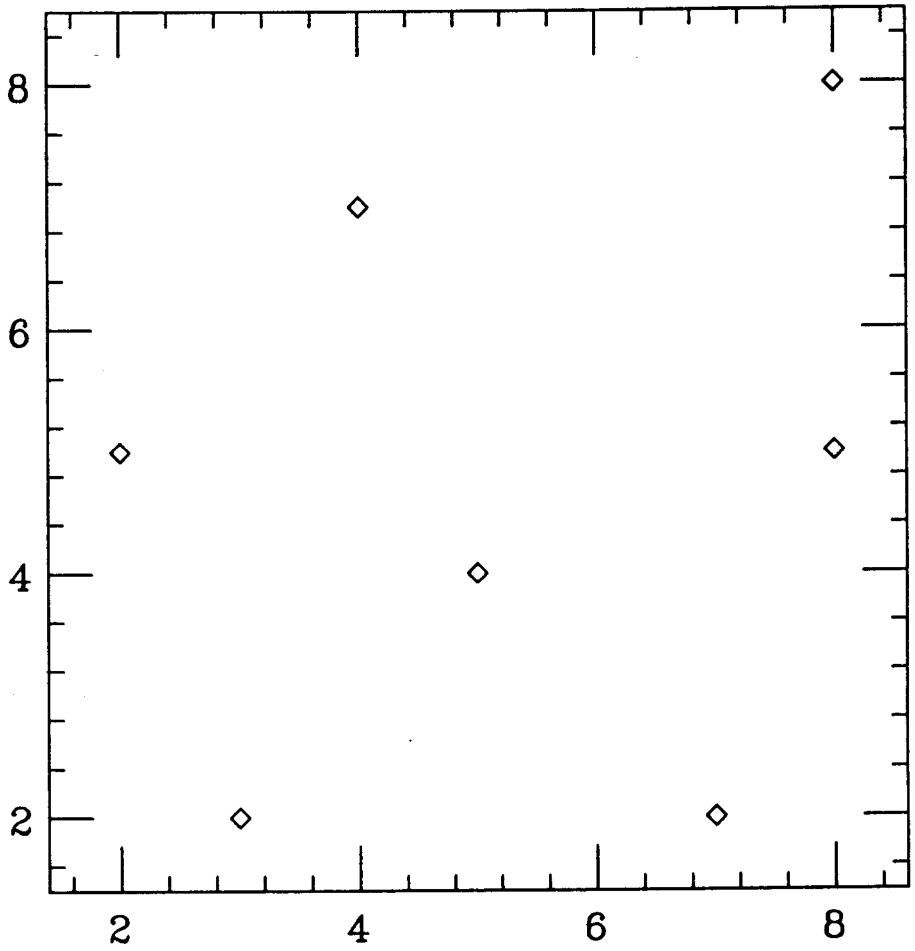
The study of the dynamics of graphs is at the beginning phase and the analysis of the subgraphs *KNN* and *KMST* seems to offer a spring of unsolved problems.

The interest in studying the dynamics of both graphs seems to be beyond that of mere speculation by the fact that they are of great interest in practical "data analysis" and "clustering" problems. The topics in this paper cover jointly discrete topology, computational geometry, and combinatorics. Several questions, regarding the cardinality of **PL** and **PS**, as function of **d** and N, are still open.

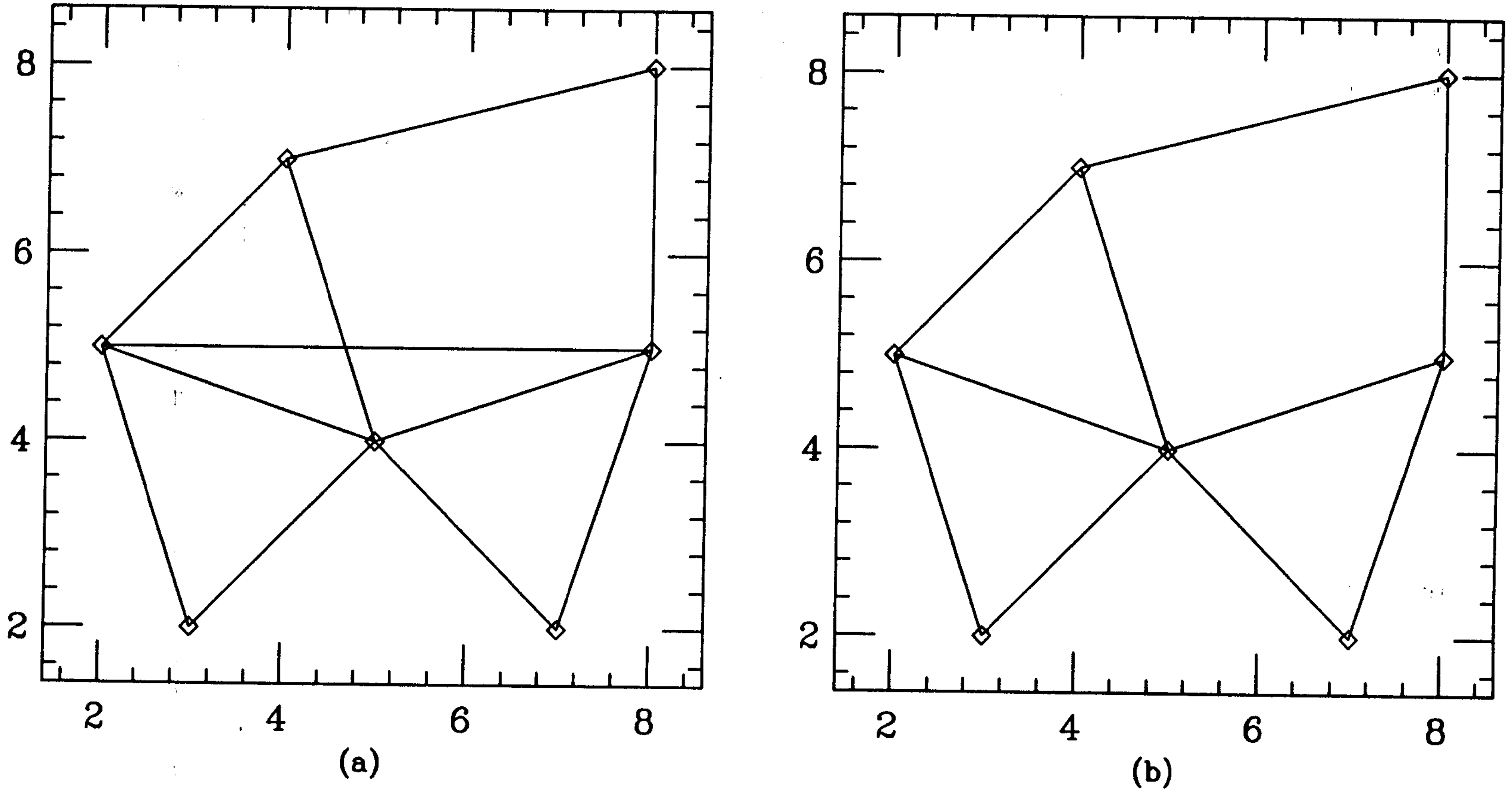
Further investigation could study the dynamics of different classes of graphs, e.g. the *t-transitive* graphs [11] seems to offer a good example in which algebraic and combinatorics problems are combined.

Table 1

d	N	PS	PL	K_{max}
		%	%	
1	5	0.	100.	3
2	»	12.5	24.	»
4	»	6.	24.	»
6	»	0.	30.	»
8	»	0.	30.	»
10	»	16.	30.	»
1	7	0.	0.	4
2	»	0.	0.	»
4	»	8.	20.	»
6	»	7.	22.	»
8	»	6.55	24.	»
10	»	6.5	32.	»
1	10	0.	0.	6
2	»	0.	0.	»
4	»	0.	12.	»
6	»	0.	22.	»
8	»	0.	20.	»
10	»	26.5	28.	»
1	20	0.	0.	11
2	»	0.	0.	»
4	»	0.	0.	»
6	»	0.	0.	»
8	»	0.	8.	»
10	»	22.	16.	»



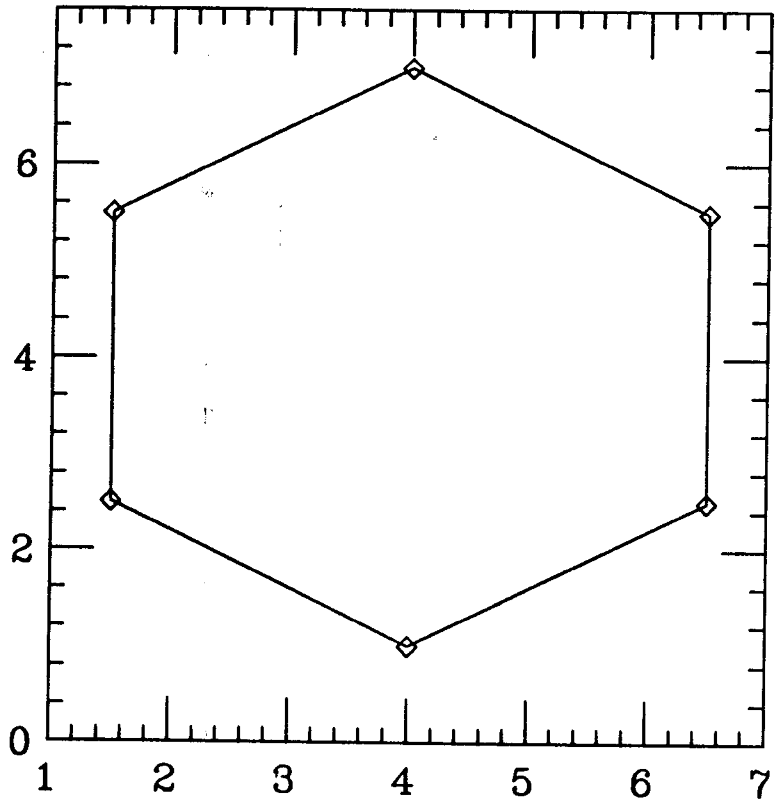
10-86 Fig.1 Graph G. 5581A1



10-86

Fig.2 (a) 2MST, (b) 2NN of the graph G.

5581A2



10-86 Fig.3 Example of PL graph

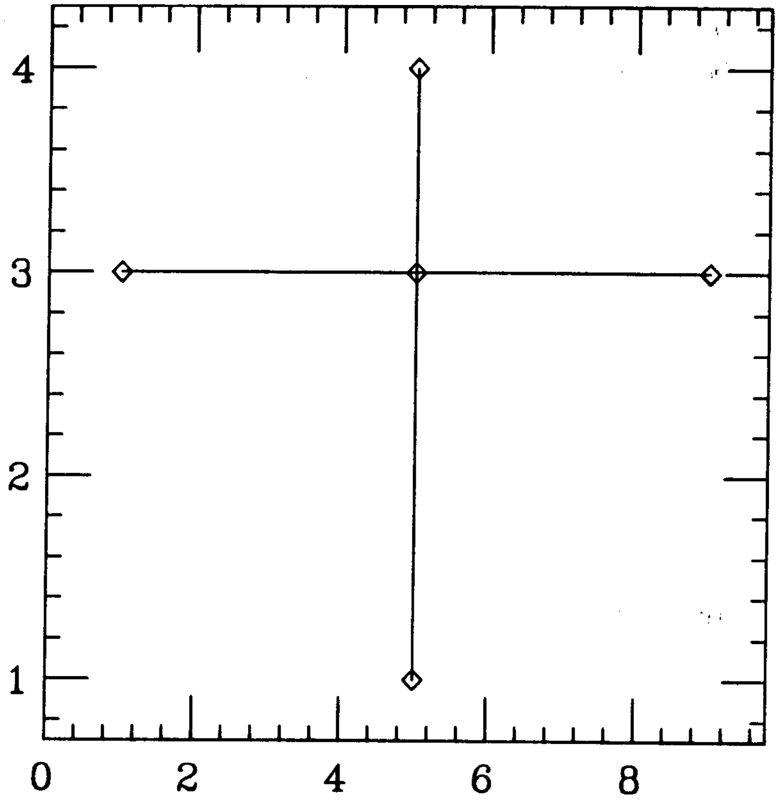


Fig.4 Example of PS graph 5581A3

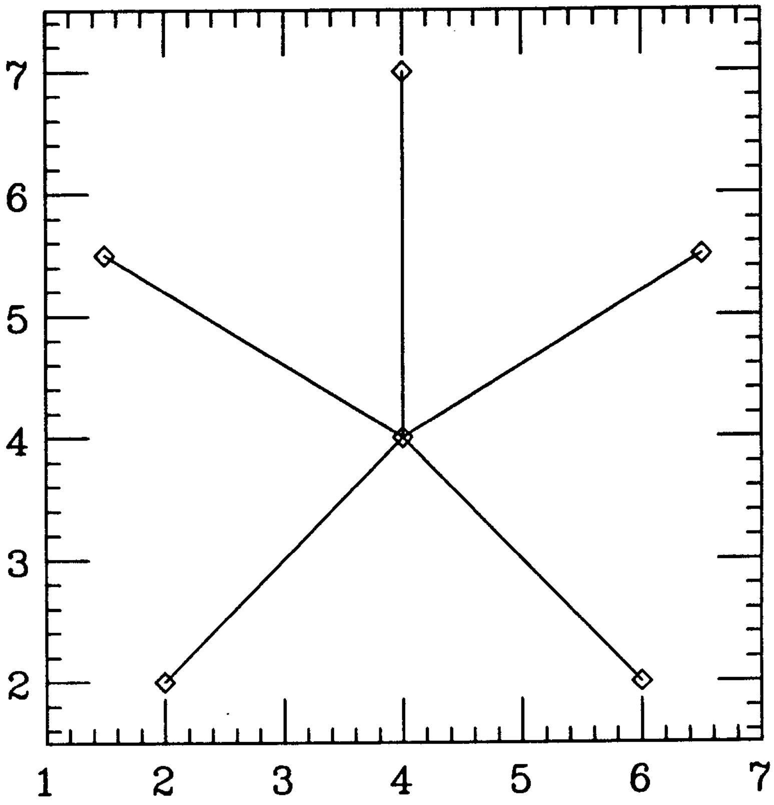
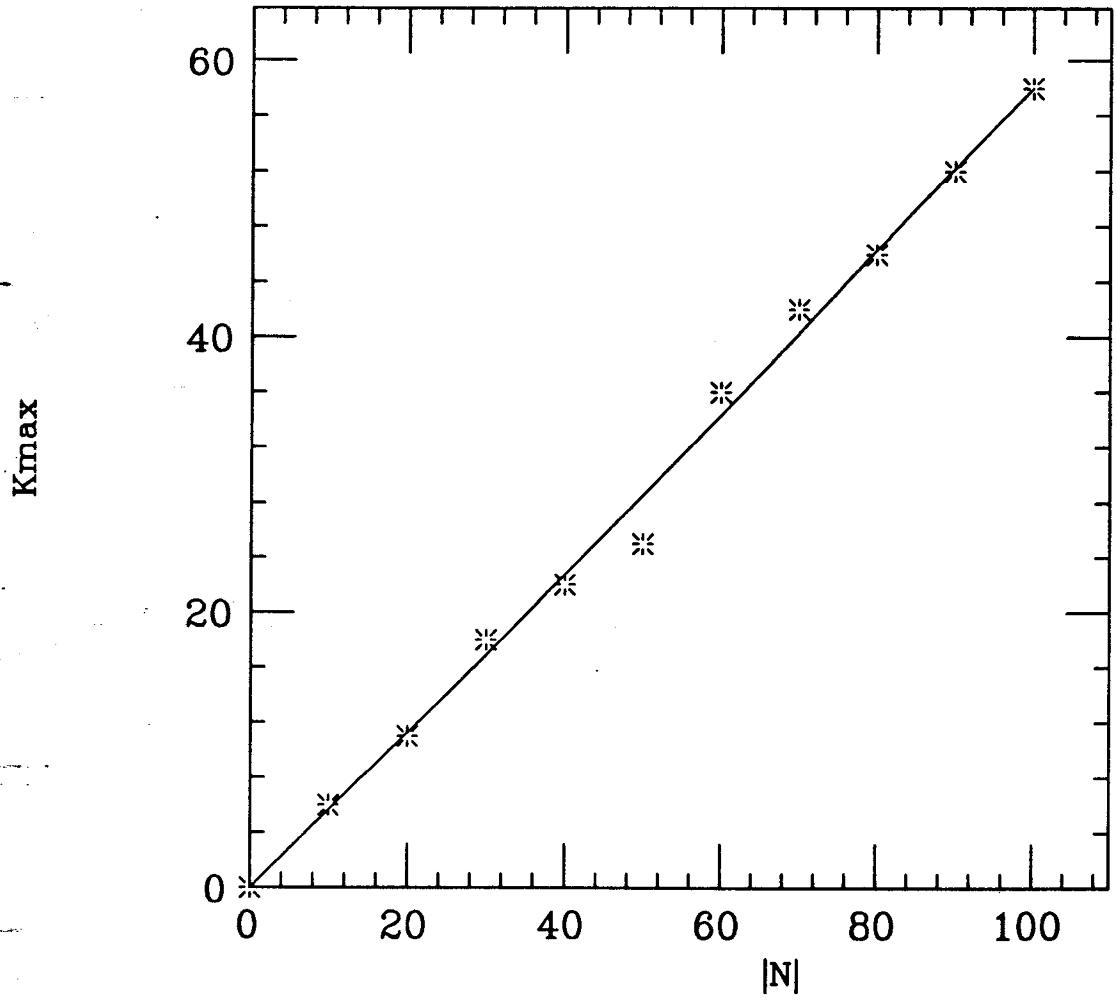


Fig.5 Example of graph with $\lceil |N| \rceil <$ Kmax $< |N|-1$

10-86 5581A4



10-86

Fig.6 Kmax Vs. |N|

5581A5