TRABAJO FIN DE MÁSTER

MÁSTER UNIVERSITARIO EN ANÁLISIS Y VISUALIZACIÓN DE DATOS MASIVOS

Caracterización de equipos informáticos mediante clustering en una red empresarial

Octubre de 2020

Javier Artiga Garijo

Dirigido por Luis Miguel Garay Gallastegui

unir LA UNIVERSIDAD EN INTERNET

unir UNIVERSIDAD INTERNACIONAL DE LA RIOJA | ESCUELA DE INGENIERÍA

INTRODUCCIÓN

INTRODUCCIÓN

- ¿Podemos clasificar en categorías relevantes las direcciones IP de una gran red empresarial, según su comportamiento de red?
- ¿Cuáles serían esas categorías?
- ¿Seremos capaces de identificar comportamientos sospechosos en base a esta clasificación?

INTRODUCCIÓN

- ¿Podemos clasificar en categorías relevantes las direcciones IP de una gran red empresarial, según su comportamiento de red?
- ¿Cuáles serían esas categorías?
- ¿Seremos capaces de identificar comportamientos sospechosos en base a esta clasificación?
- Objetivo:

"Diseñar y probar un sistema que
categorice los equipos finales de una red empresarial
y detecte comportamientos anómalos."

ESTADO DEL ARTE

- Aprendizaje automático en clasificación de tráfico
- Detección de anomalías sobre actividad de red
- Clustering

ESTADO DEL ARTE

- Aprendizaje automático en clasificación de tráfico
  - Enfoques basados en flujos
  - Enfoques basados en hosts
- Detección de anomalías sobre actividad de red

- Clustering

ESTADO DEL ARTE

- Aprendizaje automático en clasificación de tráfico
  - Enfoques basados en flujos
  - Enfoques basados en hosts
- Detección de anomalías sobre actividad de red
  - Medidas de seguridad basadas en firmas
  - Medidas de seguridad basadas en comportamientos anómalos
- Clustering

ESTADO DEL ARTE

- Aprendizaje automático en clasificación de tráfico
  - Enfoques basados en flujos
  - Enfoques basados en hosts
- Detección de anomalías sobre actividad de red
  - Medidas de seguridad basadas en firmas
  - Medidas de seguridad basadas en comportamientos anómalos
- Clustering

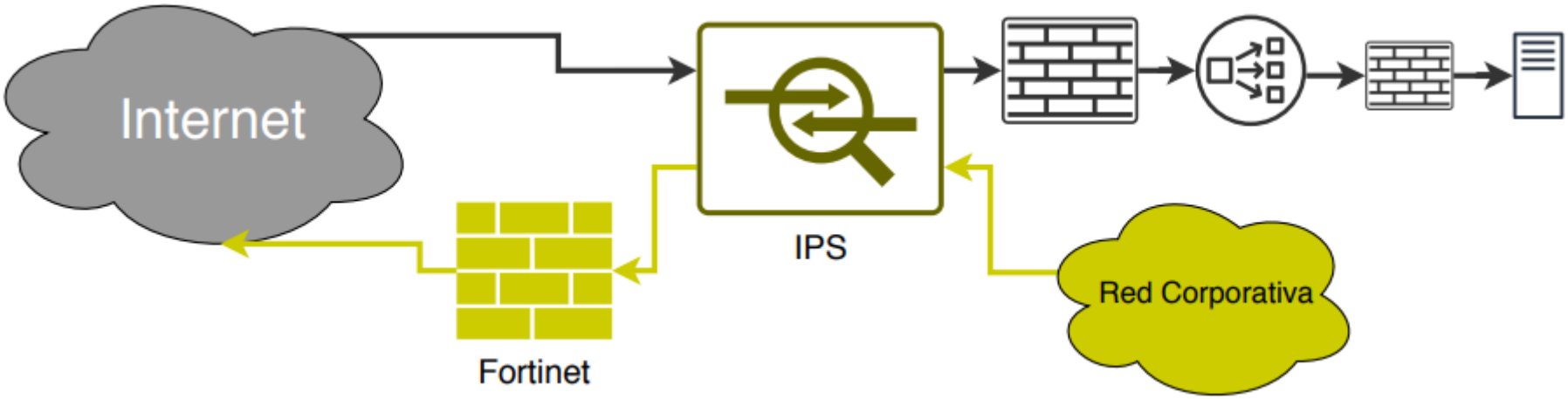
  Ventajas del aprendizaje automático: autónomo, robusto, adaptable, puede hallar patrones complejos.

  Alto coste de datos etiquetados a priori, convienen técnicas no supervisadas.

  Numerosos trabajos demuestran la utilidad del clustering para esta tarea.

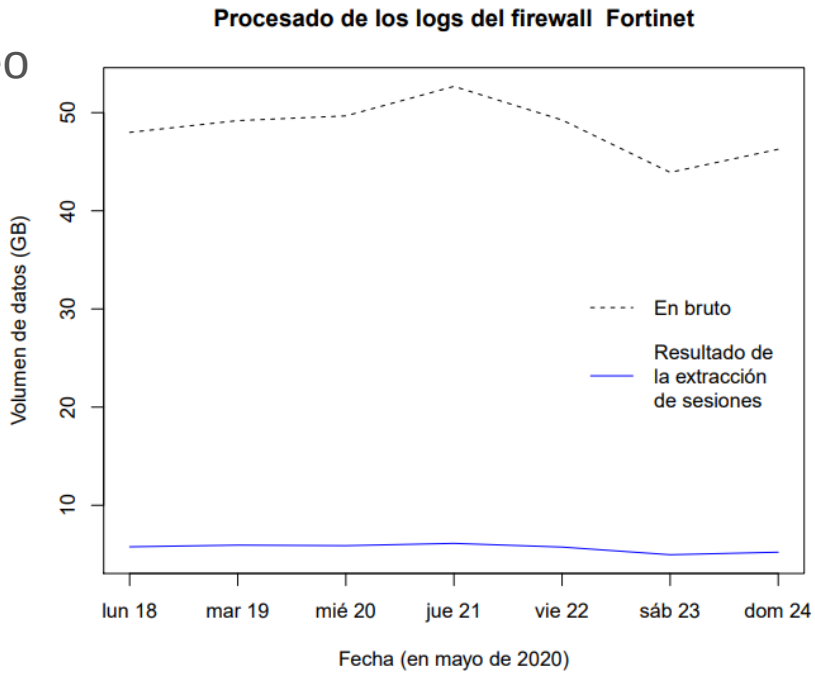
DESARROLLO DE LA CONTRIBUCIÓN

- PRESENTACIÓN DEL ESCENARIO



DESARROLLO DE LA CONTRIBUCIÓN

- EXTRACCIÓN DE SESIONES
  - Se extraen unos 20M sesiones/día.
  - Cada una se representa mediante el vector de características de la sesión.
  - Para el desarrollo se hizo un muestreo del 5% (1M sesiones por día) sobre 7 días de datos.



DESARROLLO DE LA CONTRIBUCIÓN

- PREPROCESADO PARA EL CLUSTERING
  - Consiste en resumir todas las sesiones del día para cada IP origen, aplicando agregaciones y calculando métricas.
  - Se obtienen matrices diarias.

DESARROLLO DE LA CONTRIBUCIÓN

- PREPROCESADO PARA EL CLUSTERING

Característica	Explicación
Número de IPs destino únicas	IPs distintas a las que se ha conectado una IP origen
Protocolos usados	2 si IP origen ha usado TCP y UDP, 1 si UDP, 0 si TCP
Número de puertos origen únicos	Puertos de nivel transporte usados por una IP origen
Número de puertos destino únicos	Puertos a los que se ha conectado una IP origen
Nivel de anomalía medio	Media de $N_{\text{anomalía}}$ en las sesiones de una IP origen
Nivel de amenaza medio	Media de N_{amenaza} en las sesiones de una IP origen
Prioridad máxima	Prioridad más crítica vista en eventos de una IP origen
Número de eventos	Suma total de los eventos de una IP origen
Media de la duración de sesión	Duración media de las sesiones de una IP origen
Desv. estándar de la duración de sesión	Desv. est. de la duración de sesiones de una IP origen
Nº sesiones activas en horas nocturnas	Sesiones activas de 00:00 a 08:00
Nº sesiones activas en horas de trabajo	Sesiones activas de 08:01 a 16:00
Nº ses. activas en horas después del trabajo	Sesiones activas de 16:01 a 23:59

Tabla 1: Características con las que se resumen las sesiones en matrices diarias

DESARROLLO DE LA CONTRIBUCIÓN

- SELECCIÓN DE CARACTERÍSTICAS

Característica	Explicación
Número de IPs destino únicas	IPs distintas a las que se ha conectado una IP origen
Protocolos usados	2 si IP origen ha usado TCP y UDP, 1 si UDP, 0 si TCP
Número de puertos origen únicos	Puertos de nivel transporte usados por una IP origen
Número de puertos destino únicos	Puertos a los que se ha conectado una IP origen
Nivel de anomalía medio	Media de $N_{\text{anomalía}}$ en las sesiones de una IP origen
Nivel de amenaza medio	Media de N_{amenaza} en las sesiones de una IP origen
Prioridad máxima	Prioridad más crítica vista en eventos de una IP origen
Número de eventos	Suma total de los eventos de una IP origen
Media de la duración de sesión	Duración media de las sesiones de una IP origen
Desv. estándar de la duración de sesión	Desv. est. de la duración de sesiones de una IP origen
Nº sesiones activas en horas nocturnas	Sesiones activas de 00:00 a 08:00
Nº sesiones activas en horas de trabajo	Sesiones activas de 08:01 a 16:00
Nº ses. activas en horas después del trabajo	Sesiones activas de 16:01 a 23:59

Tabla 1: Características con las que se resumen las sesiones en matrices diarias

- Son principalmente 6 características las que influyen en la clasificación.
- Las características temporales no eran tan decisivas como se creía.

DESARROLLO DE LA CONTRIBUCIÓN

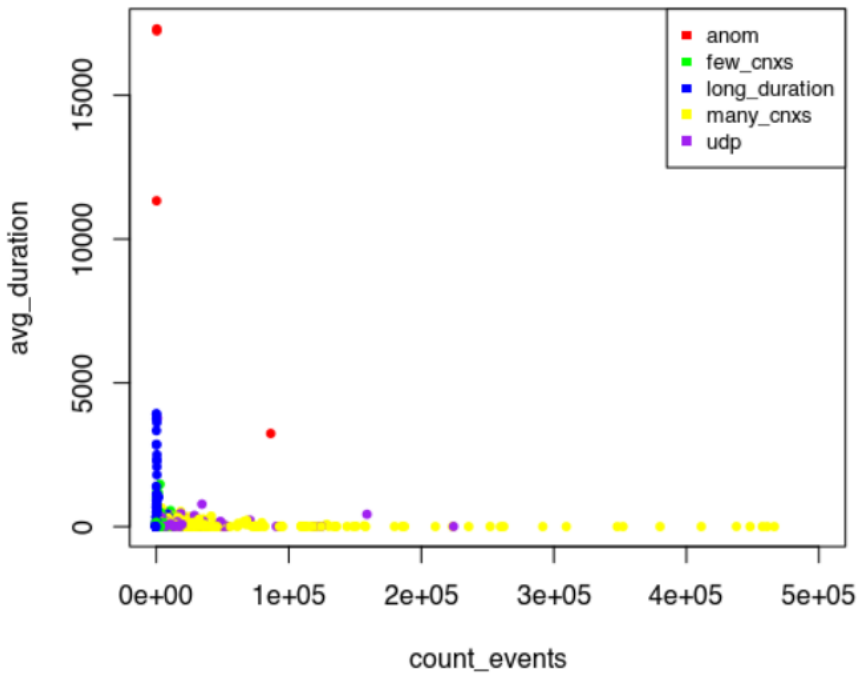
- EVALUACIONES EXPERIMENTALES
  - A través del método del codo y múltiples pruebas, se determina $k = 5$.
  - Se mide la calidad del clustering mediante:
    - Ratio de sumas de cuadrados $\left({SS_{between\ clusters}}/{SS_{total}}\right)$
    - Coeficiente de silueta

RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - Muchas conexiones
  - Pocas conexiones
  - Sesiones UDP
  - Conexiones largas
  - Anomalías

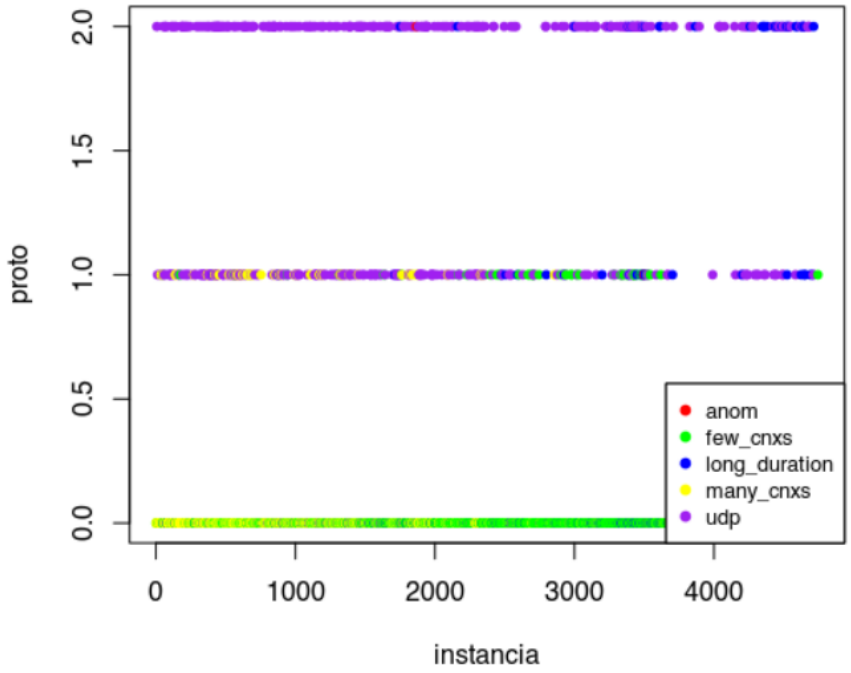
RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - Muchas conexiones
  - Pocas conexiones
  - Sesiones UDP
  - Conexiones largas
  - **Anomalías**



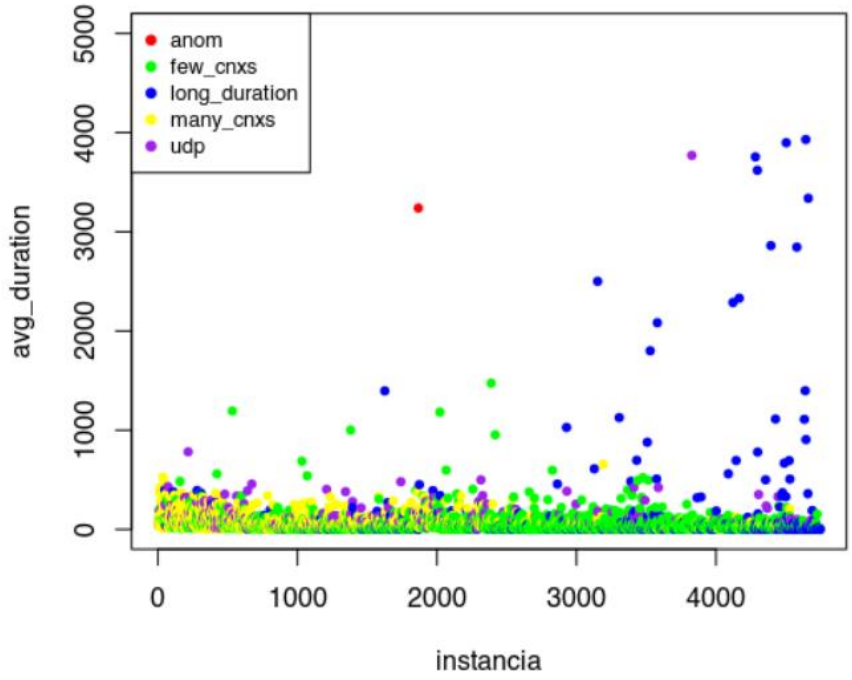
RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - Muchas conexiones
  - Pocas conexiones
  - **Sesiones UDP**
  - Conexiones largas
  - Anomalías



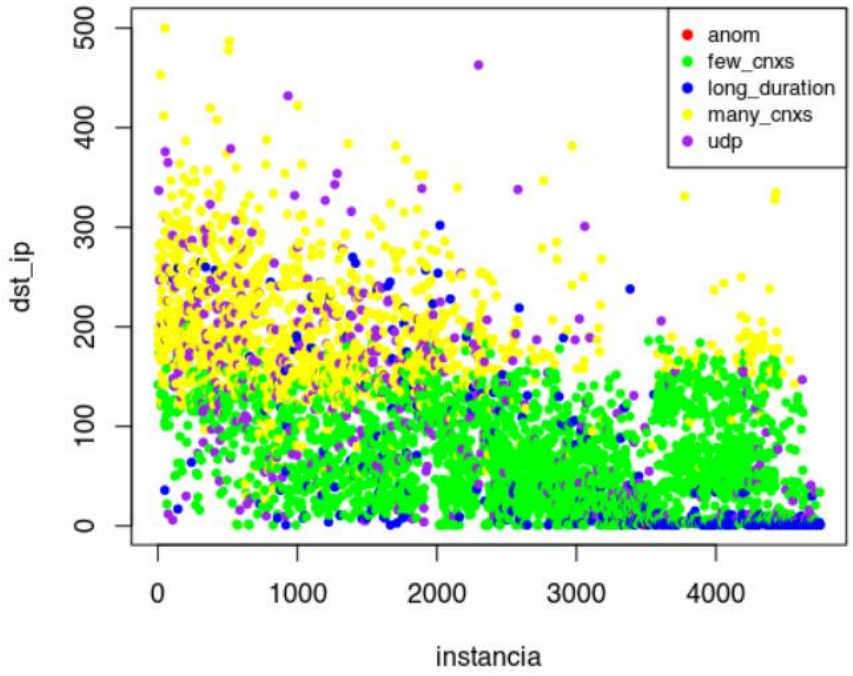
RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - Muchas conexiones
  - Pocas conexiones
  - Sesiones UDP
  - **Conexiones largas**
  - Anomalías



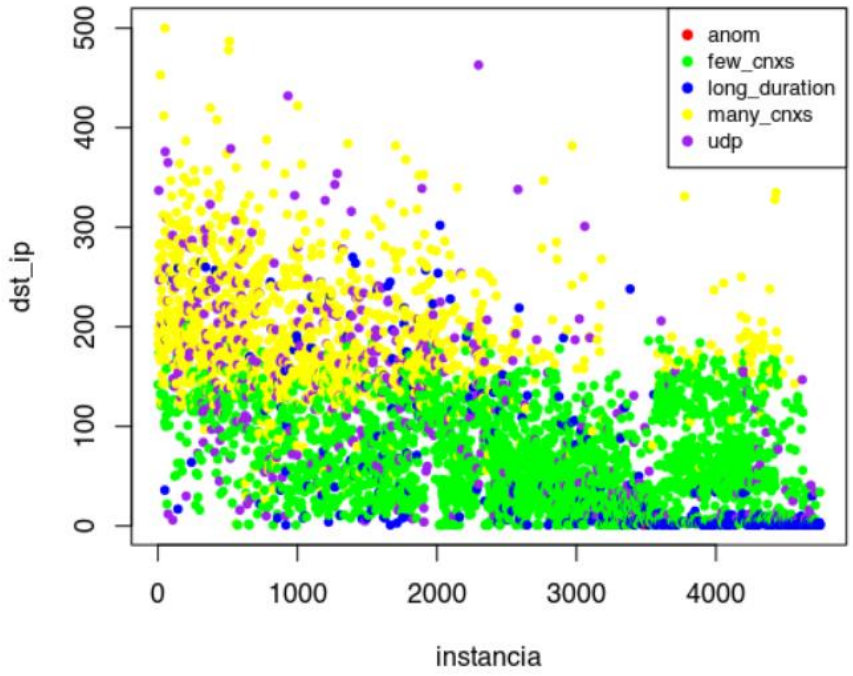
RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - **Muchas conexiones**
  - Pocas conexiones
  - Sesiones UDP
  - Conexiones largas
  - Anomalías



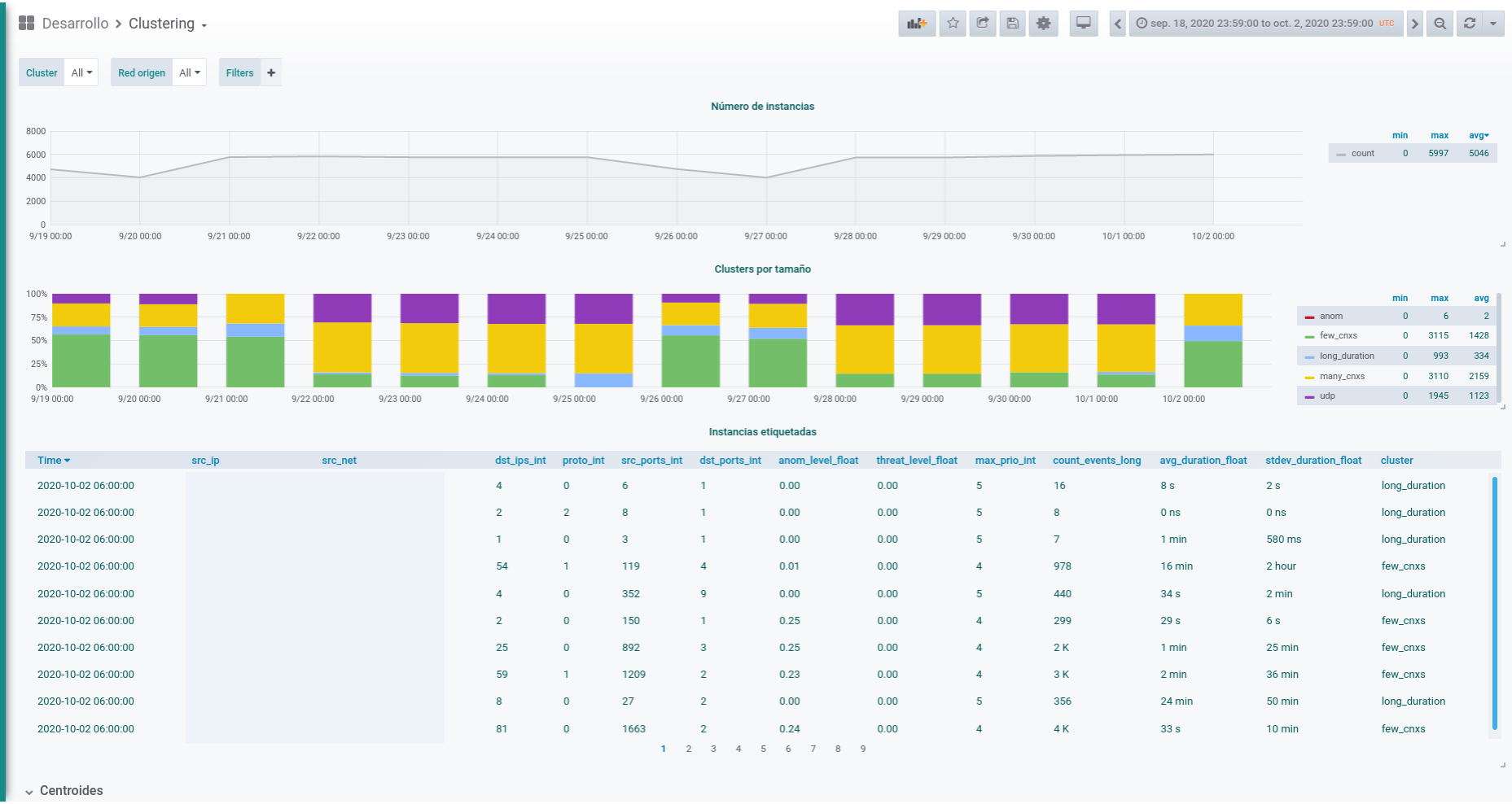
RESULTADOS

- COMPOSICIÓN DE LOS CLUSTERS:
  - Muchas conexiones
  - **Pocas conexiones**
  - Sesiones UDP
  - Conexiones largas
  - Anomalías



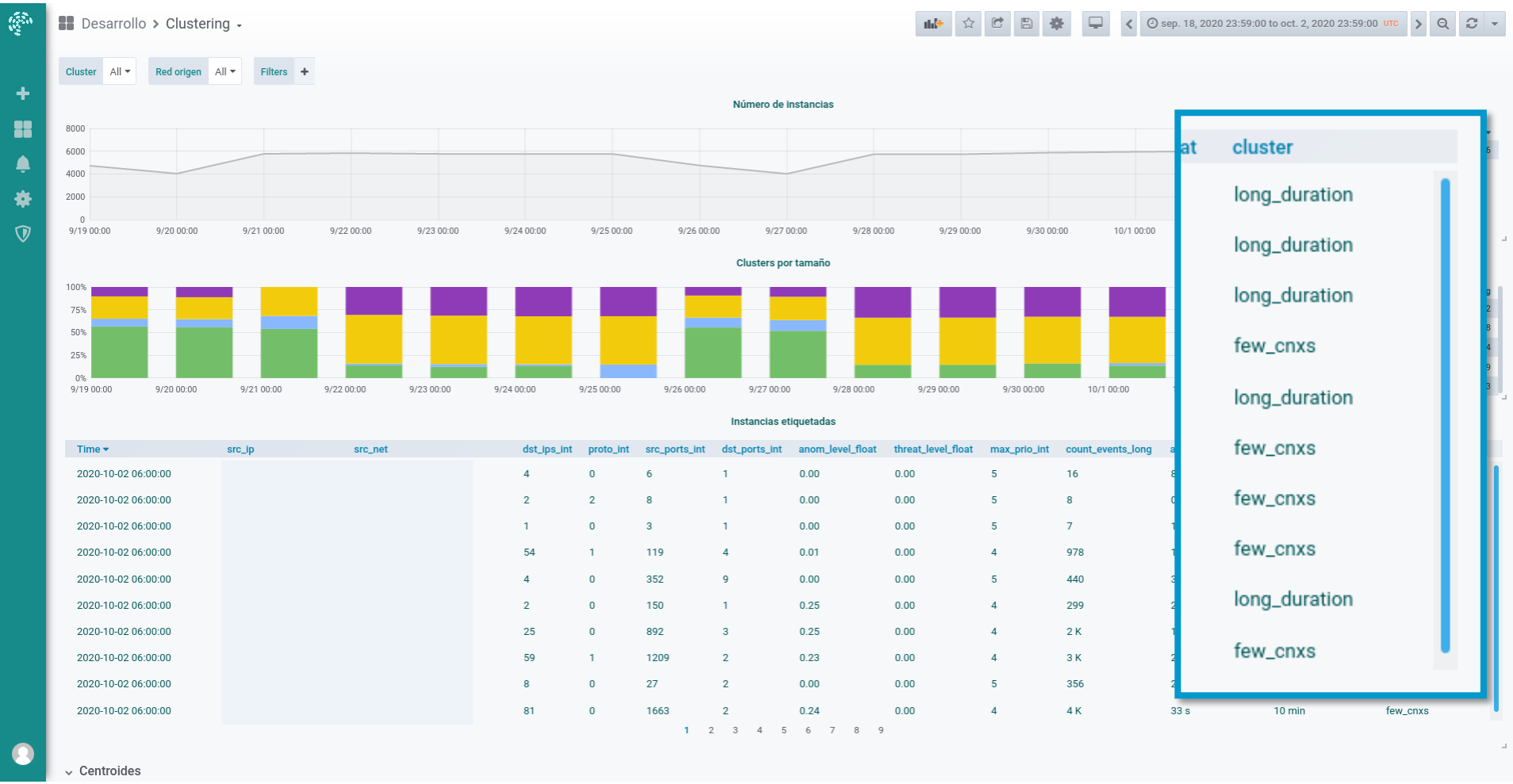
RESULTADOS

- EVALUACIÓN EN ESCENARIO REAL



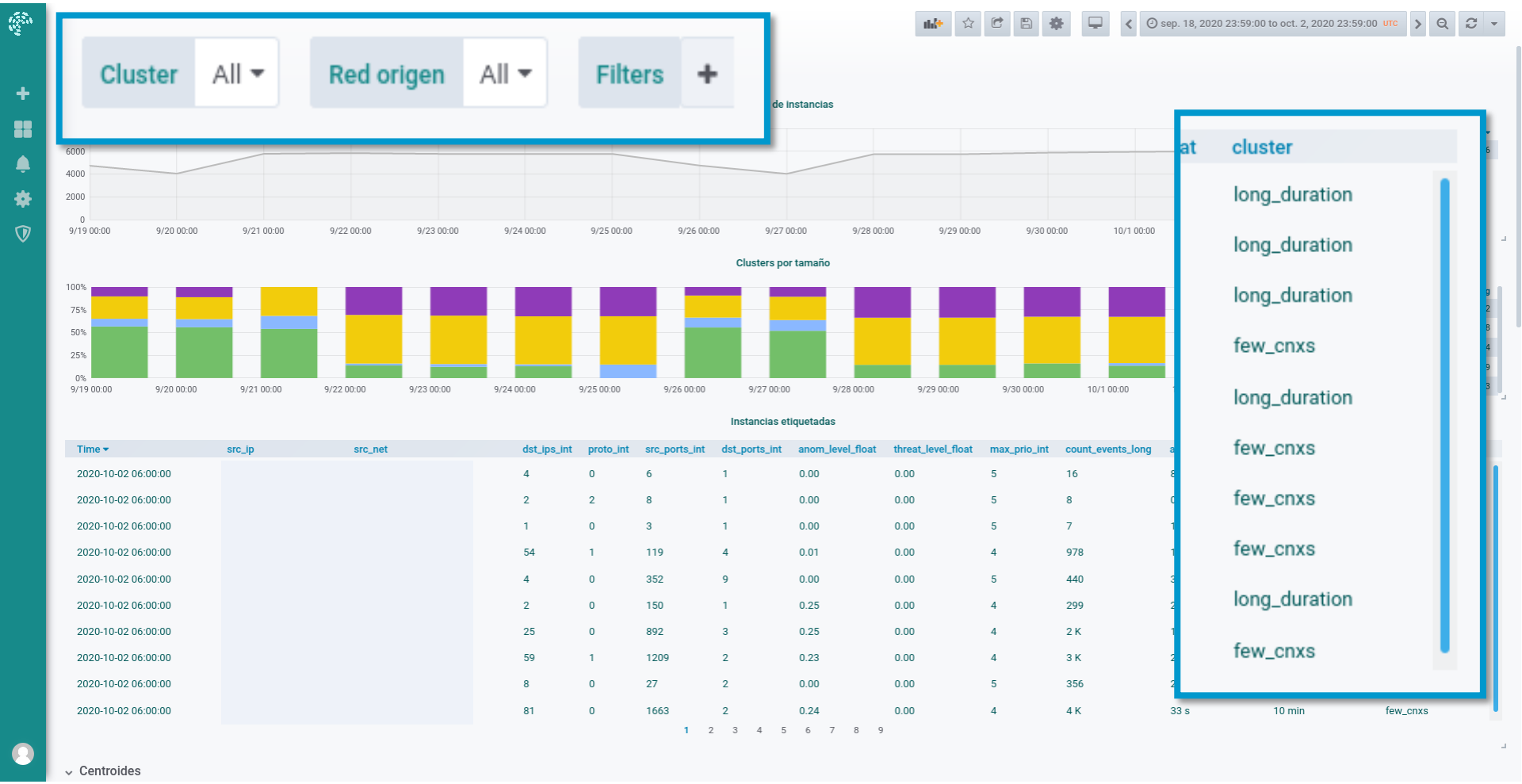
RESULTADOS

- EVALUACIÓN EN ESCENARIO REAL



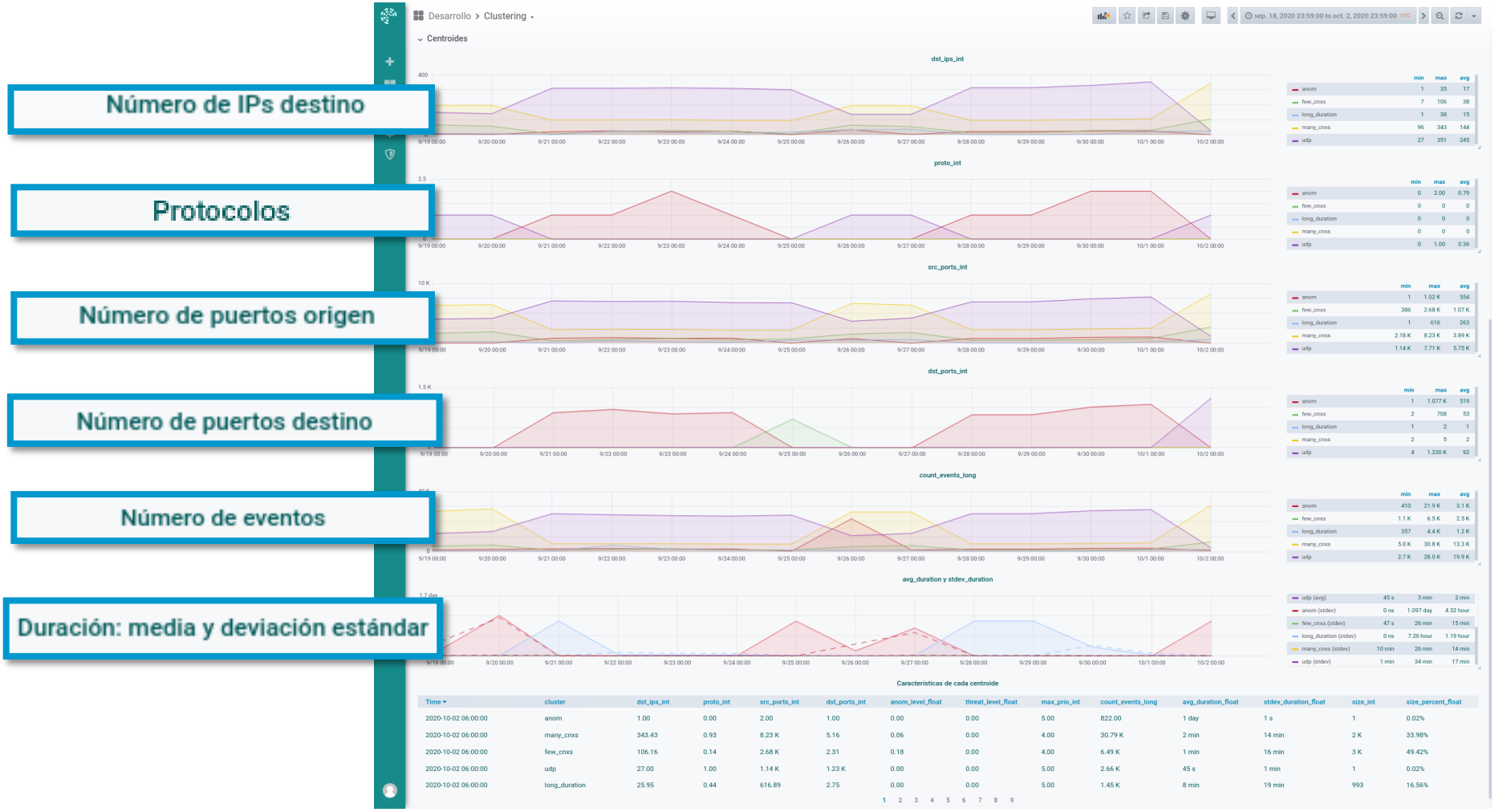
RESULTADOS

- EVALUACIÓN EN ESCENARIO REAL



RESULTADOS

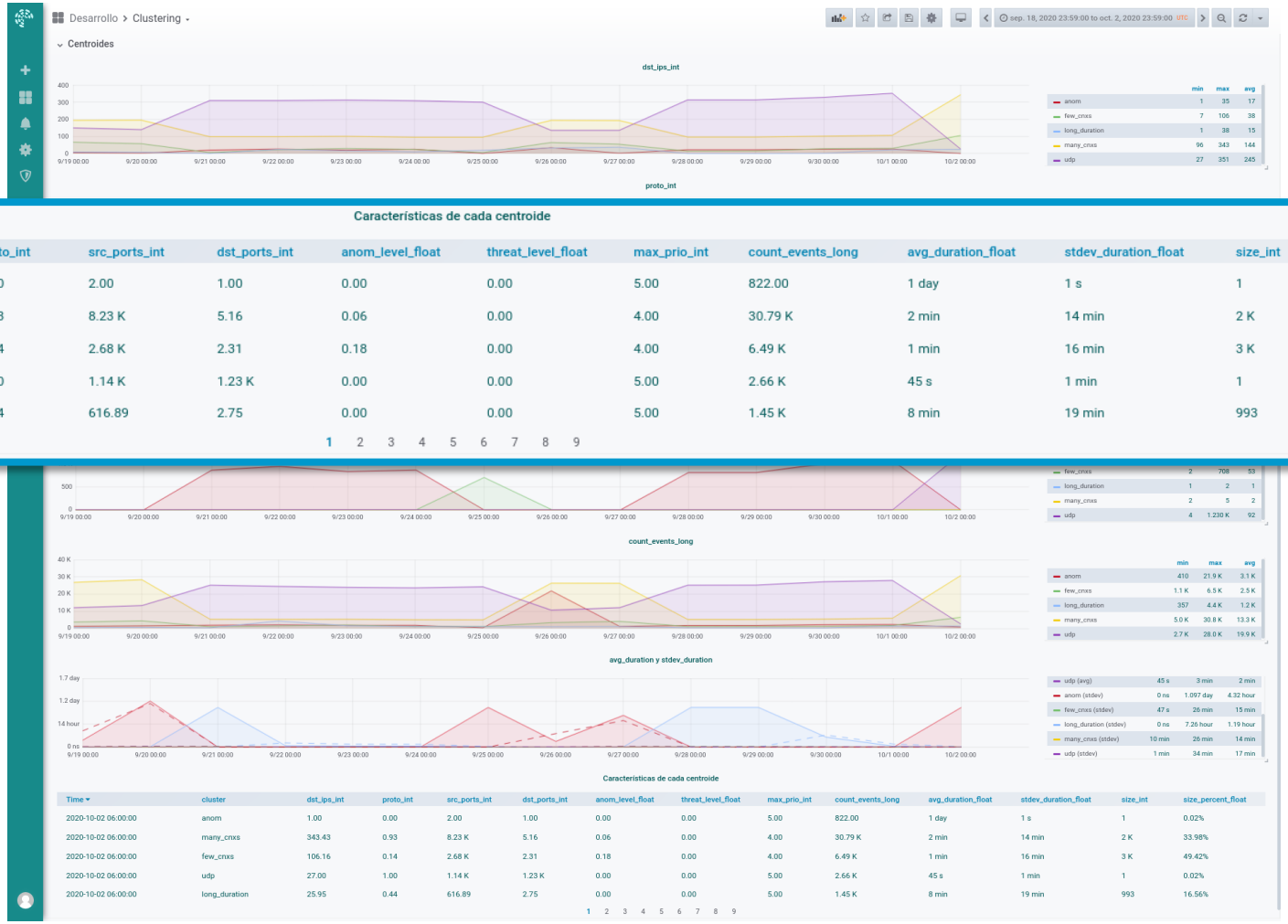
- EVALUACIÓN EN ESCENARIO REAL



RESULTADOS

- EVALUACIÓN EN ESCENARIO REAL

cluster
anom
many_cnxs
few_cnxs
udp
long_duration



Características de cada centroide

dst_ips_int	proto_int	src_ports_int	dst_ports_int	anom_level_float	threat_level_float	max_prio_int	count_events_long	avg_duration_float	stdev_duration_float	size_int
1.00	0.00	2.00	1.00	0.00	0.00	5.00	822.00	1 day	1 s	1
343.43	0.93	8.23 K	5.16	0.06	0.00	4.00	30.79 K	2 min	14 min	2 K
106.16	0.14	2.68 K	2.31	0.18	0.00	4.00	6.49 K	1 min	16 min	3 K
27.00	1.00	1.14 K	1.23 K	0.00	0.00	5.00	2.66 K	45 s	1 min	1
25.95	0.44	616.89	2.75	0.00	0.00	5.00	1.45 K	8 min	19 min	993

1 2 3 4 5 6 7 8 9

CONCLUSIONES Y LÍNEAS FUTURAS

- Podemos clasificar en categorías relevantes las direcciones IP de una gran red empresarial, según su comportamiento de red.
- La categoría “anomalías” captura comportamiento sospechosos, aunque no necesariamente malintencionados.
- Esta aportación puede tener una aplicación práctica inmediata.

CONCLUSIONES Y LÍNEAS FUTURAS

La investigación podría continuar:

- Incorporando otros firewalls como fuente de datos, e incluso correlándolos.
- Aplicando este sistema de clustering a conexiones externas.
- Incrementando la granularidad dentro de las categorías normales mediante una segunda clusterización.